

Evaluasi Komparatif *Embedding Modern* dan *Ensemble Learning* untuk Analisis Sentimen Ulasan Aplikasi *Mobile Banking*

Comparative Evaluation of Modern Embeddings and Ensemble Learning for Sentiment Analysis of Mobile Banking Application Reviews

¹I Wayan Adi Wiratama*, ²Putu Hendra Suputra, ³I Made Gede Sunarya

^{1,2,3}Program Studi Ilmu Komputer Program Pascasarjana Universitas Pendidikan Ganesha

^{1,2}Jl. Udayana No. 11, Singaraja, Kabupaten Buleleng, Bali 81116, Indonesia

*e-mail: adi.wiratama@student.undiksha.ac.id

(received: 1 September 2026, revised: 9 September 2026, accepted: 10 September 2026)

Abstrak

Ulasan pengguna aplikasi *mobile banking* menjadi sumber informasi penting untuk mengevaluasi kualitas layanan digital. Namun, representasi teks konvensional seperti *Term Frequency-Inverse Document Frequency (TF-IDF)* memiliki keterbatasan dalam menangkap konteks semantik, sedangkan ketidakseimbangan distribusi kelas dapat memengaruhi kinerja klasifikasi sentimen. Penelitian ini bertujuan mengevaluasi secara komparatif representasi teks dan algoritma klasifikasi untuk analisis sentimen ulasan aplikasi *mobile banking*. *Dataset* penelitian terdiri atas 9.885 ulasan pengguna aplikasi *Livin' by Mandiri* yang diperoleh dari *Google Play Store*. Penelitian membandingkan tiga metode representasi teks, yaitu *TF-IDF*, *Multilingual E5 Large Instruct*, dan *Nomic Embed v2*, dengan empat algoritma klasifikasi, yaitu *Naive Bayes*, *Support Vector Machine*, *XGBoost*, dan *CatBoost* sehingga menghasilkan 12 kombinasi model. *Borderline-SMOTE* diterapkan pada data latih untuk menangani ketidakseimbangan kelas, sedangkan optimasi *hyperparameter* dilakukan menggunakan *BayesSearchCV*. Evaluasi dilakukan menggunakan *accuracy*, *precision*, *recall*, *F1-score*, dan *F1-macro* melalui *stratified 5-fold cross-validation* pada data latih dan pengujian pada data uji. Hasil penelitian menunjukkan bahwa *Multilingual E5 Large Instruct* menghasilkan rata-rata *F1-macro* tertinggi sebesar 0,950. Kombinasi *CatBoost* dengan *Multilingual E5 Large Instruct* menghasilkan performa individual terbaik dengan *accuracy* sebesar 98% dan *F1-macro* sebesar 0,97. Hasil tersebut menunjukkan bahwa pemilihan kombinasi representasi teks dan algoritma klasifikasi berperan penting dalam menentukan kinerja analisis sentimen ulasan aplikasi *mobile banking* berbahasa Indonesia.

Kata kunci: *ensemble learning, mobile banking, embedding modern, oversampling, Analisis sentimen*

Abstract

User reviews of mobile banking applications provide an important source of information for evaluating the quality of digital services. However, conventional text representations such as *Term Frequency-Inverse Document Frequency (TF-IDF)* have limitations in capturing semantic context, while class imbalance can affect sentiment classification performance. This study aims to comparatively evaluate text representations and classification algorithms for sentiment analysis of mobile banking application reviews. The dataset consists of 9,885 user reviews of the *Livin' by Mandiri* application collected from the *Google Play Store*. The study compares three text representation methods, namely *TF-IDF*, *Multilingual E5 Large Instruct*, and *Nomic Embed v2*, with four classification algorithms: *Naive Bayes*, *Support Vector Machine (SVM)*, *XGBoost*, and *CatBoost*, resulting in 12 model combinations. *Borderline-SMOTE* was applied to the training data to address class imbalance, while hyperparameter optimization was performed using *BayesSearchCV*. Model performance was evaluated using *accuracy*, *precision*, *recall*, *F1-score*, and *macro-F1* through *stratified 5-fold cross-validation* on the training data and evaluation on the test set. The results show that *Multilingual E5 Large Instruct* achieved the highest average *macro-F1* score of 0.950. The

<http://sistemasi.ftik.unisi.ac.id>

combination of CatBoost and Multilingual E5 Large Instruct achieved the best individual performance, with an accuracy of 98% and a macro-F1 score of 0.97. These findings demonstrate that the choice of text representation and classification algorithm plays an important role in determining the performance of sentiment analysis for Indonesian-language mobile banking application reviews.

Keywords: ensemble learning, mobile banking, modern embeddings, oversampling, Sentiment analysis

1 Pendahuluan

Transformasi digital telah mendorong industri perbankan untuk menghadirkan layanan yang semakin mudah, cepat, dan fleksibel melalui aplikasi *mobile banking*. Seiring dengan meningkatnya penggunaan aplikasi tersebut, pengguna secara aktif memberikan ulasan dan penilaian melalui platform distribusi aplikasi seperti *Google Play Store*. Ulasan pengguna tidak hanya mencerminkan tingkat kepuasan terhadap layanan yang diberikan, tetapi juga memuat informasi mengenai kualitas fitur, kemudahan penggunaan, performa aplikasi, serta berbagai permasalahan teknis yang dialami pengguna. Oleh karena itu, analisis sentimen terhadap ulasan pengguna menjadi salah satu pendekatan yang efektif untuk memperoleh informasi secara otomatis dari data tekstual dalam jumlah besar sehingga dapat dimanfaatkan sebagai dasar evaluasi dan pengembangan layanan digital [1]. Meskipun telah banyak diterapkan, analisis sentimen pada ulasan aplikasi berbahasa Indonesia masih menghadapi tantangan akibat karakteristik ulasan yang umumnya berupa teks pendek, tidak terstruktur, dan menggunakan bahasa informal. Representasi teks konvensional seperti *Term Frequency-Inverse Document Frequency (TF-IDF)* belum mampu menangkap hubungan semantik maupun konteks antar kata, sedangkan *embedding modern* berbasis *Transformer* menghasilkan representasi semantik yang lebih kaya pada berbagai tugas *Natural Language Processing* multibahasa, termasuk Bahasa Indonesia [2] [3]. Selain itu, ketidakseimbangan distribusi kelas pada *dataset* ulasan juga dapat menurunkan kemampuan model dalam mengenali kelas minoritas. Teknik *oversampling*, khususnya *Borderline-SMOTE*, dapat digunakan untuk membentuk sampel sintetis pada area *decision boundary* sehingga membantu meningkatkan representasi kelas minoritas [4].

Berbagai penelitian terdahulu mengenai analisis sentimen ulasan aplikasi berbahasa Indonesia masih banyak menggunakan *TF-IDF* yang dipadukan dengan algoritma klasifikasi konvensional seperti *Naive Bayes* dan *Support Vector Machine (SVM)* [5][6]. Meskipun pendekatan tersebut mampu menghasilkan performa klasifikasi yang baik, kemampuannya dalam mengenali kelas minoritas masih dapat terpengaruh ketika berhadapan dengan *dataset* yang tidak seimbang [7]. Penelitian yang lebih baru mulai memanfaatkan *embedding* berbasis *Transformer* dan teknik *oversampling* untuk mengatasi keterbatasan tersebut. Namun, penelitian-penelitian tersebut belum memberikan gambaran yang komprehensif mengenai bagaimana perbedaan representasi teks, mulai dari metode konvensional hingga *embedding modern*, memengaruhi performa berbagai algoritma klasifikasi pada kondisi *dataset* yang tidak seimbang. Celah tersebut menjadi penting karena peningkatan kualitas representasi teks tidak selalu menghasilkan performa klasifikasi yang sama pada setiap algoritma. Oleh karena itu, diperlukan evaluasi dalam kerangka eksperimen yang sama agar perbedaan performa dapat dibandingkan secara lebih objektif. Dalam konteks analisis sentimen ulasan aplikasi *mobile banking* berbahasa Indonesia, kajian yang secara sistematis membandingkan *TF-IDF*, *embedding modern Multilingual E5 Large Instruct* dan *Nomic Embed v2*, serta algoritma *Naive Bayes*, *SVM*, *XGBoost*, dan *CatBoost* masih terbatas. Selain itu, kombinasi representasi teks tersebut dengan penanganan *class imbalance* menggunakan *Borderline-SMOTE* dan optimasi *hyperparameter* berbasis *Bayesian Optimization* belum dievaluasi secara terpadu dalam satu rancangan eksperimen.

Berdasarkan celah tersebut, penelitian ini melakukan evaluasi komparatif terhadap tiga metode representasi teks dan empat algoritma klasifikasi dalam kerangka eksperimen yang sama, sehingga menghasilkan 12 kombinasi model yang dapat dibandingkan secara langsung. Seluruh kombinasi dievaluasi pada *dataset* ulasan aplikasi *mobile banking* berbahasa Indonesia menggunakan *accuracy*, *precision*, *recall*, *F1-score*, dan *F1-macro* [8]. Kontribusi penelitian ini adalah (1) menyediakan *benchmark* terkontrol yang mempertemukan *TF-IDF*, *Multilingual E5 Large Instruct*, dan *Nomic Embed v2* dengan *Naive Bayes*, *SVM*, *XGBoost*, dan *CatBoost*; (2) mendokumentasikan pengaruh interaksi antara representasi teks dan algoritma klasifikasi pada data ulasan berbahasa Indonesia yang

tidak seimbang; dan (3) mengidentifikasi kombinasi terbaik pada korpus yang diuji. Dengan demikian, kebaruan penelitian diposisikan sebagai evaluasi empiris terpadu pada korpus ulasan Livin' by Mandiri, bukan sebagai klaim bahwa satu metode berlaku universal untuk seluruh aplikasi *mobile banking*.

2 Tinjauan Literatur

2.1 Penelitian Terdahulu dan Analisis Sentimen

Penelitian mengenai analisis sentimen ulasan aplikasi berkembang seiring meningkatnya penggunaan layanan digital dan jumlah umpan balik pengguna pada platform distribusi aplikasi. Ulasan pengguna dapat digunakan untuk mengidentifikasi persepsi terhadap kualitas layanan, fitur, kemudahan penggunaan, serta permasalahan yang dialami. Sejumlah penelitian terdahulu pada analisis sentimen berbahasa Indonesia menggunakan pendekatan *machine learning* dengan representasi teks konvensional seperti *Term Frequency-Inverse Document Frequency (TF-IDF)*, yang kemudian dikombinasikan dengan algoritma seperti *Naive Bayes* dan *Support Vector Machine (SVM)* [6] [9]. Pendekatan tersebut mampu menghasilkan performa klasifikasi yang baik, tetapi masih memiliki keterbatasan dalam merepresentasikan hubungan semantik dan konteks antar kata, terutama pada ulasan yang bersifat pendek, informal, dan tidak terstruktur.

Perkembangan penelitian menunjukkan adanya pergeseran menuju penggunaan representasi berbasis *Transformer* untuk memperoleh informasi semantik yang lebih kaya. *Multilingual E5 Large Instruct* merupakan salah satu *embedding* multibahasa yang dikembangkan untuk menghasilkan representasi semantik pada berbagai bahasa [2]. Kondisi tersebut menunjukkan bahwa pemilihan representasi teks merupakan aspek penting dalam menentukan performa klasifikasi sentimen dan perlu dievaluasi pada konteks data yang spesifik, termasuk ulasan aplikasi *mobile banking* berbahasa Indonesia.

2.2 Dataset Penelitian

Objek penelitian ini adalah *dataset* ulasan pengguna aplikasi Livin' by Mandiri yang diperoleh dari *Google Play Store* melalui *dataset* yang dipublikasikan di Kaggle pada Juli 2025. *Dataset* terdiri atas 9.885 ulasan berbahasa Indonesia yang telah diberi label sentimen positif dan negatif. Distribusi data menunjukkan kondisi *class imbalance*, dengan sekitar 81,5% ulasan berlabel negatif dan 18,5% berlabel positif. *Dataset* dibagi secara terstratifikasi dengan rasio 70:30 menggunakan *random_state=42*, sehingga diperoleh 6.919 data latih dan 2.966 data uji. Rasio 70:30 dipilih karena memberikan proporsi yang seimbang antara ketersediaan data untuk pelatihan dan jumlah data yang memadai untuk pengujian, serta didukung oleh penelitian terdahulu yang menunjukkan kinerja yang baik pada rasio tersebut [10]. Setiap kombinasi dilatih menggunakan data latih dan dioptimalkan melalui *hyperparameter tuning* berbasis *stratified 5-fold cross-validation* pada data latih, kemudian dievaluasi pada data uji yang tidak digunakan selama proses pelatihan maupun optimasi.

2.3 Representasi Teks dan Embedding

Perkembangan representasi teks dalam analisis sentimen menunjukkan pergeseran dari pendekatan konvensional seperti *TF-IDF* menuju representasi berbasis *Transformer* yang mampu menangkap informasi semantik dan konteks secara lebih baik. Pendekatan berbasis *Transformer* telah diterapkan dalam analisis sentimen berbahasa Indonesia, antara lain melalui *fine-tuning IndoBERT* untuk klasifikasi sentimen pada berita pariwisata Indonesia [11]. Pada pendekatan tersebut, model *pre-trained Transformer* disesuaikan (*fine-tuning*) dengan data dan tugas klasifikasi tertentu. Berbeda dengan pendekatan *fine-tuning*, penelitian ini memanfaatkan model *modern embedding* yang telah dilatih sebelumnya sebagai ekstraktor representasi fitur tanpa melakukan penyesuaian parameter model terhadap *dataset* penelitian. Pendekatan ini memungkinkan representasi semantik dari teks diperoleh dalam bentuk vektor yang selanjutnya digunakan sebagai masukan bagi algoritma klasifikasi. Dua model *modern embedding* yang digunakan adalah *Multilingual E5 Large Instruct* dan *Nomic Embed v2*, yang dipilih untuk mengevaluasi kemampuan representasi teks multibahasa, khususnya pada ulasan berbahasa Indonesia.

Multilingual E5 Large Instruct merupakan keluarga model *embedding* multilingual yang dikembangkan untuk menghasilkan representasi semantik teks melalui pendekatan *contrastive learning* [2]. Sementara itu, *Nomic Embed v2* merupakan model *text embedding* multilingual berbasis arsitektur *Sparse Mixture of Experts (MoE)* yang dirancang untuk menghasilkan representasi teks dengan efisiensi komputasi yang lebih baik dibandingkan pendekatan *dense* berukuran besar [12]. Meskipun model *embedding modern* menunjukkan kemampuan yang tinggi dalam merepresentasikan informasi semantik, performanya tidak selalu konsisten pada seluruh jenis tugas. Hasil *benchmark MTEB* menunjukkan bahwa tidak terdapat satu metode *text embedding* yang secara konsisten mendominasi seluruh tugas evaluasi, sehingga pemilihan representasi teks perlu mempertimbangkan karakteristik *dataset* dan tujuan klasifikasi [3]. Penelitian ini menempatkan *TF-IDF* sebagai *baseline* dan membandingkannya dengan *Multilingual E5 Large Instruct* serta *Nomic Embed v2* dalam prosedur eksperimen yang sama.

Konfigurasi *embedding* pada penelitian ini adalah *Multilingual E5 Large Instruct* menggunakan model *intfloat/multilingual-e5-large-instruct*, awalan *query*, dan menghasilkan vektor berdimensi 1.024. *Nomic Embed v2* menggunakan model *nomic-ai/nomic-embed-text-v2-moe* dengan *trust_remote_code=True*, *prompt passage*, dan menghasilkan vektor berdimensi 768. Kedua model merupakan model *pre-trained* yang digunakan untuk inferensi fitur, dan tidak dilakukan *fine-tuning* pada penelitian ini. Panjang urutan maksimum yang digunakan mengikuti konfigurasi *SentenceTransformers* model, yaitu 512 *token*.

2.4 Algoritma Klasifikasi dan Ensemble Learning

Selain representasi teks, algoritma klasifikasi merupakan faktor yang dapat memengaruhi performa analisis sentimen. *Naive Bayes* dan *SVM* banyak digunakan dalam klasifikasi teks karena mampu menangani data dengan jumlah fitur yang tinggi [13] [6]. Perkembangan selanjutnya menunjukkan penggunaan algoritma *ensemble learning* seperti *XGBoost* dan *CatBoost* yang menggunakan pendekatan *boosting* untuk mempelajari pola kompleks pada data. Penggunaan algoritma tersebut menjadi semakin relevan ketika fitur yang digunakan berupa *dense embedding* karena karakteristik representasinya berbeda dengan fitur *sparse* yang dihasilkan *TF-IDF*. Hasil penelitian terdahulu menunjukkan bahwa performa algoritma klasifikasi dapat berbeda ketika diterapkan pada representasi teks yang berbeda. Oleh karena itu, peningkatan kualitas representasi teks tidak secara otomatis menghasilkan performa klasifikasi yang sama pada seluruh algoritma. Kondisi tersebut menjadi dasar bagi penelitian ini untuk membandingkan *Naive Bayes*, *SVM*, *XGBoost*, dan *CatBoost* pada tiga metode representasi teks sehingga setiap algoritma klasifikasi dapat dievaluasi secara konsisten dalam 12 kombinasi model sebagaimana terlihat pada Tabel 1.

Tabel 1 Kombinasi metode representasi teks dan algoritma klasifikasi

Representasi Teks	Algoritma Klasifikasi
TF-IDF	Naive Bayes
	SVM
	XGBoost
	CatBoost
Multilingual E5 Large Instruct	Naive Bayes
	SVM
	XGBoost
	CatBoost
Nomic Embed v2	Naive Bayes
	SVM
	XGBoost
	CatBoost

2.5 Preprocessing

Tahapan *preprocessing* dalam penelitian analisis sentimen pada data berbahasa Indonesia umumnya mencakup *case folding*, *tokenizing*, *stopword removal*, dan *stemming* untuk menyeragamkan bentuk teks serta mengurangi *noise* pada data [14]. Dalam penelitian ini, khusus

untuk representasi *TF-IDF*, dilakukan tahapan *case folding*, pembersihan karakter, normalisasi spasi, *stopword removal*, *stemming* Bahasa Indonesia, *negation chaining*, serta penghapusan data kosong setelah *preprocessing* [15] [16]. Vektorisasi menggunakan *TfidfVectorizer(ngram_range=(1, 2))*, yang menghasilkan 35.446 fitur pada data yang diproses. Untuk *XGBoost* dan *CatBoost*, fitur *sparse* direduksi menggunakan *TruncatedSVD (LSA)* untuk memperoleh representasi berdimensi lebih rendah sebelum digunakan dalam proses klasifikasi. Pada *embedding modern*, teks dipertahankan sedekat mungkin dengan bentuk aslinya agar konteks semantik, termasuk bahasa informal, *slang*, dan singkatan, tetap tersedia.

2.6 Penyeimbangan Data dan Pelatihan Model

Ketidakseimbangan kelas merupakan salah satu permasalahan penting dalam klasifikasi karena algoritma pembelajaran dapat cenderung memprioritaskan kelas mayoritas dan mengabaikan kelas minoritas. Berbagai pendekatan telah dikembangkan untuk mengatasi permasalahan tersebut, antara lain melalui *data-level approaches*, *algorithm-level approaches*, dan *ensemble learning* [7]. *Borderline-SMOTE* diterapkan hanya pada data latih setelah pembagian *dataset*. Teknik ini membentuk sampel sintesis dari data minoritas yang berada di sekitar *decision boundary* [4]. Penelitian terdahulu pada analisis sentimen juga menunjukkan bahwa penerapan *oversampling* dapat meningkatkan performa klasifikasi pada *dataset* yang tidak seimbang, termasuk melalui penggunaan *Borderline-SMOTE* [17]. Sementara data uji tidak mengalami *oversampling* untuk mencegah *data leakage*. Optimasi *hyperparameter* diperlukan karena konfigurasi *hyperparameter* dapat memengaruhi kompleksitas, perilaku, dan performa model. Dalam penelitian ini, optimasi dilakukan menggunakan pendekatan *Bayesian Optimization* melalui *BayesSearchCV* dengan *n_iter*=20, skor *F1-macro*, *stratified 5-fold cross-validation*, dan *random_state*=42 untuk memperoleh konfigurasi *hyperparameter* yang optimal berdasarkan evaluasi *cross-validation* [18] [19]. Ruang pencarian *hyperparameter* ditetapkan berdasarkan konfigurasi eksperimen penelitian. Untuk *Naive Bayes*, parameter *alpha* dicari pada rentang 0,001–1,0 dengan distribusi *log-uniform*. Untuk *SVM*, parameter *C* dicari pada rentang 0,001–100 dengan distribusi *log-uniform*, *max_iter* juga dioptimasi pada konfigurasi *TF-IDF*. Untuk *XGBoost*, ruang pencarian meliputi *n_estimators* 100-300, *max_depth* 3-6, *learning_rate* 0,01-0,3, *subsample* 0,7-1,0, *colsample_bytree* 0,7-1,0, serta jumlah komponen *LSA* 100-200. Untuk *CatBoost*, ruang pencarian meliputi *iterations* 100-300, *depth* 4-6, dan *learning_rate* 0,05-0,2.

2.7 Lingkungan Eksperimen

Untuk memastikan proses eksperimen dapat direplikasi, penelitian ini menggunakan lingkungan komputasi yang mendukung proses ekstraksi *embedding* dan pelatihan model. Spesifikasi lingkungan eksperimen yang digunakan dalam penelitian ini disajikan pada Tabel 2.

Tabel 2 Spesifikasi lingkungan eksperimen

Komponen	Spesifikasi
Platform	Google Colaboratory
Backend	Python 3 Google Compute Engine (GPU)
GPU	NVIDIA Tesla T4
GPU Memory	15GB
RAM	12.7GB
Disk Storage	112GB

2.8 Evaluasi

Performa model dievaluasi menggunakan *accuracy*, *precision*, *recall*, *F1-score*, dan *F1-macro*. *F1-macro* digunakan sebagai salah satu metrik utama karena menghitung rata-rata *F1-score* pada setiap kelas dengan bobot yang sama, sehingga kontribusi kelas minoritas tetap diperhitungkan dalam evaluasi. Pemilihan metrik ini penting pada *dataset* dengan distribusi kelas yang tidak seimbang karena performa model dapat terlihat berbeda bergantung pada metrik evaluasi yang digunakan. Evaluasi dilakukan melalui *k-fold cross-validation* pada data latih dan pengujian akhir pada data uji

[20] [21]. Selain itu, dilakukan analisis *confusion matrix*, perbandingan performa data latih dan data uji, serta analisis *training curve* pada *XGBoost* dan *CatBoost* untuk menilai kemampuan generalisasi dan indikasi *overfitting*.

2.9 Research Gap dan Pengembangan Hipotesis

Hasil analisis penelitian terdahulu menunjukkan bahwa analisis sentimen ulasan aplikasi berbahasa Indonesia masih banyak menggunakan *TF-IDF* dan algoritma klasifikasi konvensional, sementara penggunaan *embedding* berbasis *Transformer* mulai berkembang. Namun, perbandingan *TF-IDF* dengan beberapa *embedding modern* yang dikombinasikan dengan berbagai algoritma klasifikasi dalam kondisi eksperimen yang sama masih terbatas, khususnya pada *dataset* dengan *class imbalance*. Penelitian ini mengisi celah tersebut melalui evaluasi 12 kombinasi *TF-IDF*, *Multilingual E5 Large Instruct*, dan *Nomic Embed v2* dengan *Naive Bayes*, *SVM*, *XGBoost*, dan *CatBoost*, menggunakan *Borderline-SMOTE* pada data latih dan optimasi *hyperparameter* dengan *BayesSearchCV*. Kebaruan penelitian terletak pada evaluasi terpadu berbagai representasi teks dan algoritma klasifikasi untuk analisis sentimen ulasan *mobile banking* berbahasa Indonesia. Berdasarkan hasil analisis literatur, penelitian ini mengajukan hipotesis bahwa representasi *embedding modern* menghasilkan performa klasifikasi yang lebih tinggi dibandingkan *TF-IDF* pada kombinasi model tertentu, *Multilingual E5 Large Instruct* menghasilkan performa klasifikasi yang lebih tinggi dibandingkan *Nomic Embed v2*, dan *ensemble learning* dapat menghasilkan performa klasifikasi yang kompetitif ketika dipadukan dengan representasi *embedding modern*.

3 Hasil dan Pembahasan

Penelitian ini mengevaluasi 12 kombinasi model yang merupakan hasil penggabungan 3 metode representasi teks, yaitu *TF-IDF*, *Multilingual E5 Large Instruct*, dan *Nomic Embed v2*, dengan 4 algoritma klasifikasi, yaitu *Naive Bayes*, *Support Vector Machine (SVM)*, *XGBoost*, dan *CatBoost*. Setiap kombinasi dilatih menggunakan data latih dan dioptimalkan melalui *hyperparameter tuning* berbasis *stratified 5-fold cross-validation* pada data latih, kemudian dievaluasi pada data uji yang tidak digunakan selama proses pelatihan maupun optimasi. Evaluasi dilakukan menggunakan metrik *accuracy*, *precision*, *recall*, *F1-score*, dan *F1-macro* sebagai metrik utama karena *dataset* memiliki distribusi kelas yang tidak seimbang. Hasil eksperimen menunjukkan bahwa *embedding modern* berbasis *Transformer*, khususnya *Multilingual E5 Large Instruct*, menghasilkan performa yang lebih baik dibandingkan metode *baseline TF-IDF*, sedangkan *XGBoost* dan *CatBoost* mampu menghasilkan performa tinggi ketika dipadukan dengan *embedding modern*, khususnya *Multilingual E5 Large Instruct*. Namun, performa masing-masing algoritma klasifikasi menunjukkan variasi bergantung pada representasi teks yang digunakan.

3.1 Hasil Evaluasi Representasi Teks

Hasil pengujian menunjukkan bahwa *Multilingual E5 Large Instruct* menghasilkan performa terbaik dibandingkan metode representasi teks lainnya. Perbandingan nilai *F1-macro* untuk ketiga metode representasi teks pada seluruh algoritma klasifikasi ditunjukkan pada Tabel 3. Berdasarkan Tabel 3, *Multilingual E5 Large Instruct* memperoleh rata-rata *F1-macro* sebesar 0,950, lebih tinggi dibandingkan *Nomic Embed v2* dan *TF-IDF* yang sama-sama memperoleh rata-rata sebesar 0,915.

Tabel 3 Perbandingan *F1-macro* antar metode representasi teks

Metode Embedding	Naive Bayes	SVM	XGBoost	CatBoost	Rata-rata
TF-IDF	0,92	0,93	0,91	0,90	0,915
Multilingual E5 Large Instruct	0,91	0,96	0,96	0,97	0,950
Nomic Embed v2	0,84	0,94	0,94	0,94	0,915

Performa tinggi yang dihasilkan *Multilingual E5 Large Instruct* menunjukkan bahwa *embedding* berbasis *Transformer* mampu merepresentasikan makna semantik dan hubungan kontekstual antar kata secara lebih efektif dibandingkan pendekatan berbasis frekuensi kata. Kemampuan ini menjadi penting pada *dataset* ulasan aplikasi yang memiliki karakteristik bahasa informal, *slang*, singkatan, dan struktur kalimat yang tidak baku. *Nomic Embed v2* juga menunjukkan performa yang kompetitif,

khususnya ketika dipadukan dengan algoritma *ensemble learning*. Namun, performanya masih berada di bawah *Multilingual E5 Large Instruct* pada sebagian besar skenario eksperimen. Sementara itu, *TF-IDF* menunjukkan performa yang lebih rendah karena representasi *sparse* berbasis frekuensi kata memiliki keterbatasan dalam memahami konteks dan makna semantik teks.

3.2 Hasil Evaluasi Algoritma Klasifikasi

Berdasarkan hasil eksperimen, performa algoritma klasifikasi menunjukkan variasi ketika dipadukan dengan representasi teks yang berbeda. Berdasarkan rata-rata *F1-macro* pada ketiga metode representasi teks, *Support Vector Machine (SVM)* memperoleh nilai rata-rata tertinggi sebesar 0,943, diikuti oleh *XGBoost* dan *CatBoost* yang masing-masing memperoleh nilai 0,937. Meskipun demikian, *CatBoost* menghasilkan nilai *F1-macro* individual tertinggi sebesar 0,97 ketika dipadukan dengan *Multilingual E5 Large Instruct*. Perbandingan *F1-macro* antar algoritma klasifikasi pada ketiga metode representasi teks disajikan pada Tabel 4.

Tabel 4 Perbandingan *F1-macro* antar algoritma klasifikasi

Algoritma	TF-IDF	Multilingual E5 Large Instruct	Nomic Embed v2	Rata-rata
Naive Bayes	0,92	0,91	0,84	0,890
SVM	0,93	0,96	0,94	0,943
XGBoost	0,91	0,96	0,94	0,937
CatBoost	0,90	0,97	0,94	0,937

Hasil tersebut menunjukkan bahwa algoritma *ensemble learning* memiliki kemampuan yang baik ketika dipadukan dengan *dense embedding*, terutama pada *Multilingual E5 Large Instruct*. *XGBoost* dan *CatBoost* masing-masing memperoleh nilai *F1-macro* sebesar 0,96 dan 0,97 pada representasi tersebut. Sementara itu, *Naive Bayes* menunjukkan penurunan performa yang lebih besar pada *Nomic Embed v2*, dengan nilai *F1-macro* sebesar 0,84. SVM menunjukkan performa yang relatif konsisten pada ketiga representasi teks dengan nilai *F1-macro* antara 0,93 hingga 0,96. Hasil ini menunjukkan bahwa performa algoritma klasifikasi tidak hanya dipengaruhi oleh algoritma klasifikasi, tetapi juga oleh karakteristik representasi fitur yang digunakan.

3.3 Kombinasi Model Terbaik

Dari seluruh kombinasi eksperimen yang diuji, kombinasi *Multilingual E5 Large Instruct* dan *CatBoost* menghasilkan performa terbaik dengan nilai *F1-macro* sebesar 0,97 dan *accuracy* sebesar 98%. Ringkasan peringkat performa seluruh 12 kombinasi model berdasarkan nilai *F1-macro* ditampilkan pada Tabel 5.

Tabel 5 Ringkasan performa seluruh 12 kombinasi berdasarkan *F1-macro*

Peringkat	Kombinasi	Accuracy	F1-macro
1	CatBoost + Multilingual E5 Large Instruct	98%	0,97
2	SVM + Multilingual E5 Large Instruct	98%	0,96
2	XGBoost + Multilingual E5 Large Instruct	98%	0,96
4	SVM + Nomic Embed v2	96%	0,94
4	XGBoost + Nomic Embed v2	96%	0,94
4	CatBoost + Nomic Embed v2	96%	0,94
7	SVM + TF-IDF	96%	0,93
8	Naive Bayes + TF-IDF	95%	0,92
9	Naive Bayes + Multilingual E5 Large Instruct	94%	0,91
10	XGBoost + TF-IDF	95%	0,91
11	CatBoost + TF-IDF	94%	0,90
12	Naive Bayes + Nomic Embed v2	88%	0,84

Kombinasi tersebut menghasilkan nilai *F1-macro* tertinggi, yang menunjukkan performa klasifikasi yang relatif seimbang antar kelas sentimen. Tingginya performa kombinasi tersebut menunjukkan bahwa *embedding modern* berbasis *Transformer* yang dipadukan dengan *ensemble learning* mampu menghasilkan representasi fitur yang informatif untuk mendukung klasifikasi teks

<http://sistemasi.ftik.unisi.ac.id>

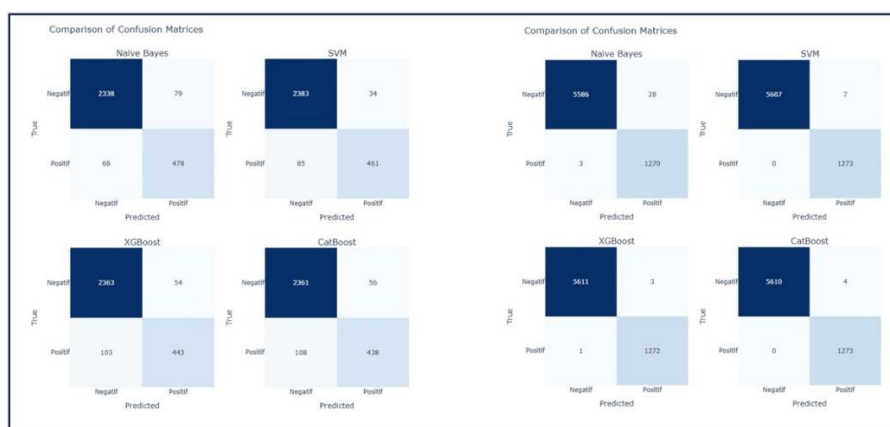
berbahasa Indonesia. Selain itu, penerapan *Borderline-SMOTE* pada data latih ditujukan untuk menangani ketidakseimbangan kelas dengan menghasilkan sampel sintetis pada area kelas minoritas yang berada di sekitar batas keputusan. Penerapan teknik tersebut mendukung proses pembelajaran model dalam mengenali distribusi kelas yang tidak seimbang, sehingga evaluasi performa pada masing-masing kelas dapat dilakukan secara lebih seimbang.

3.4 Evaluasi Generalisasi Model

Evaluasi generalisasi dilakukan untuk mengetahui kemampuan model dalam mempertahankan performanya pada data yang tidak digunakan selama proses pelatihan. Evaluasi dilakukan dengan membandingkan performa pada data latih dan data uji, serta mengamati *training curve* pada model *XGBoost* dan *CatBoost*. Hasil evaluasi generalisasi untuk masing-masing kelompok representasi teks ditunjukkan pada Gambar 1 sampai Gambar 9. Perbandingan tersebut digunakan untuk mengidentifikasi kemungkinan *overfitting* dan melihat kestabilan performa model selama proses pelatihan.

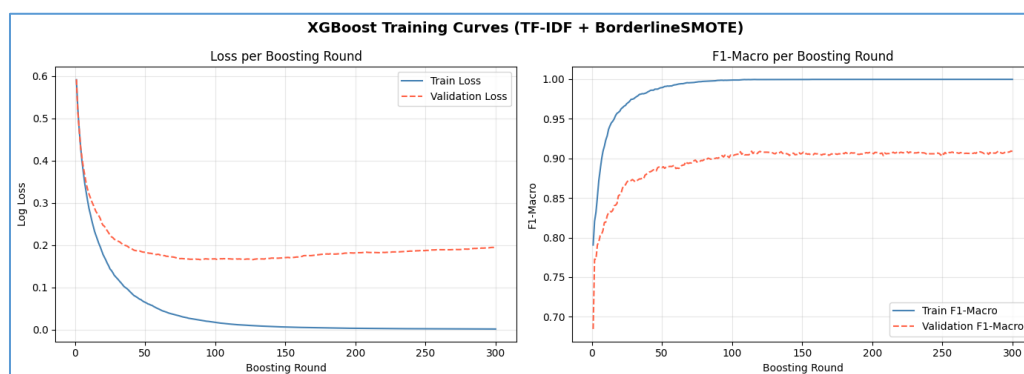
a. Eksperimen Kelompok *TF-IDF*

Perbandingan performa model pada kelompok *TF-IDF* antara data latih dan data uji dilakukan untuk mengevaluasi konsistensi kinerja model pada data yang digunakan selama pembelajaran dan data yang digunakan untuk pengujian. Selisih performa yang relatif kecil antara kedua kelompok data menunjukkan bahwa model berbasis *TF-IDF* tidak mengalami perbedaan kinerja yang ekstrem. Hasil perbandingan tersebut disajikan pada Gambar 1.



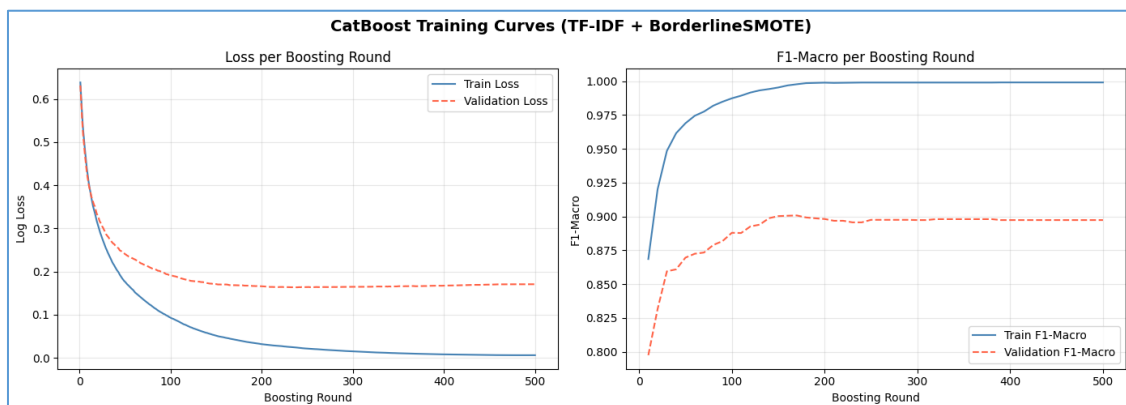
Gambar 1 Hasil perbandingan Kelompok *TF-IDF* pada data uji dan data latih

Kurva pelatihan model *XGBoost* dengan representasi *TF-IDF* digunakan untuk mengevaluasi perubahan performa selama proses pembelajaran serta membandingkan kecenderungan performa pada data pelatihan dan validasi setelah penerapan *Borderline-SMOTE* pada data latih. Kestabilan kurva validasi tanpa penurunan yang tajam menunjukkan bahwa proses pembelajaran berlangsung relatif stabil dan tidak menunjukkan indikasi *overfitting* yang kuat. Hasil kurva pelatihan tersebut disajikan pada Gambar 2.



Gambar 2 Kurva pelatihan *XGBoost* pada *TF-IDF* dengan *Borderline-SMOTE*

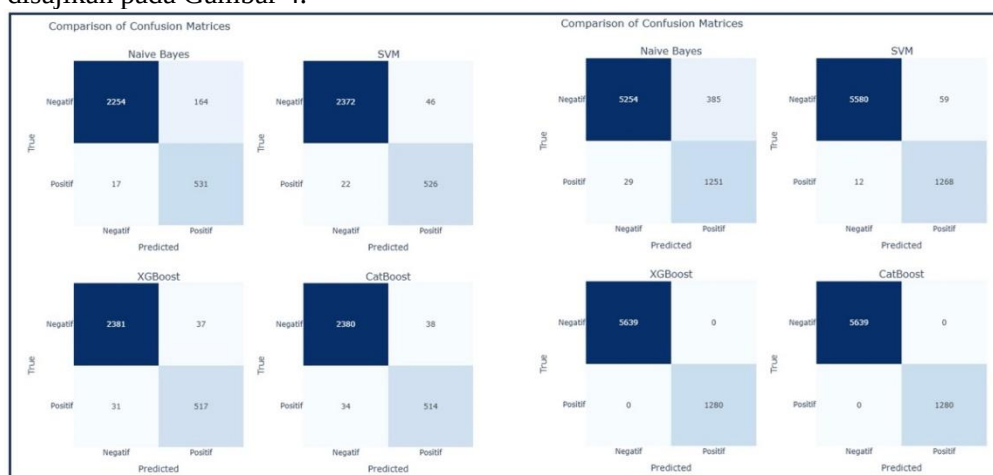
Kurva pelatihan model *CatBoost* dengan representasi *TF-IDF* digunakan untuk menganalisis perkembangan kinerja model selama proses pelatihan setelah penerapan *Borderline-SMOTE* pada data latih. Performa validasi yang relatif stabil terhadap performa pelatihan menunjukkan bahwa proses pembelajaran berlangsung secara konsisten dan tidak menunjukkan indikasi *overfitting* yang kuat. Hasil analisis kurva pelatihan tersebut disajikan pada Gambar 3.



Gambar 3 Kurva pelatihan *CatBoost* pada *TF-IDF* dengan *Borderline-SMOTE*

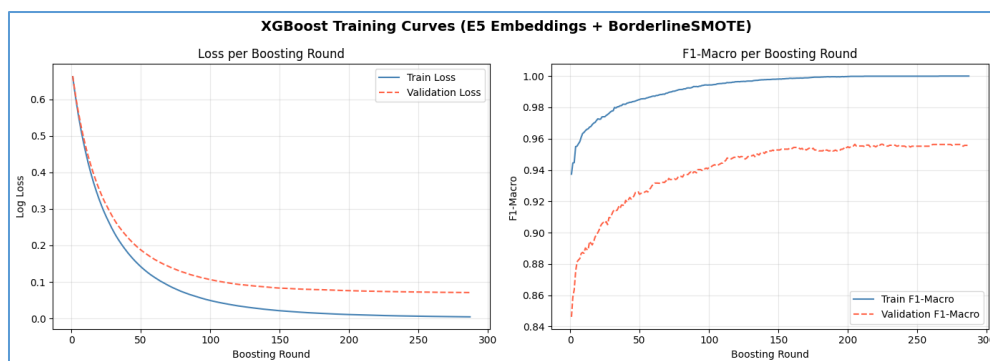
b. Eksperimen Kelompok *Multilingual E5 Large Instruct*

Perbandingan performa model pada kelompok *Multilingual E5 Large Instruct* antara data latih dan data uji dilakukan untuk mengevaluasi konsistensi kinerja model ketika diterapkan pada data yang tidak digunakan selama proses pelatihan. Selisih performa yang relatif kecil antara data latih dan data uji menunjukkan bahwa model berbasis *Multilingual E5 Large Instruct* mampu mempertahankan kinerja pada data yang tidak digunakan selama proses pelatihan. Hasil perbandingan performa tersebut disajikan pada Gambar 4.



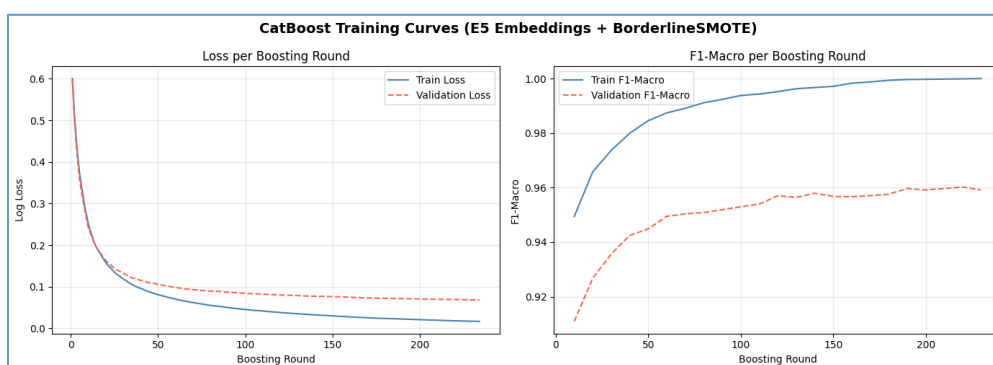
Gambar 4 Hasil perbandingan kelompok *multilingual e5 large instruct* pada data uji dan data latih

Kurva pelatihan model *XGBoost* dengan representasi *Multilingual E5 Large Instruct* digunakan untuk menganalisis perkembangan kinerja model selama proses pelatihan setelah penerapan *Borderline-SMOTE* pada data latih. Performa validasi yang relatif stabil terhadap performa pelatihan menunjukkan bahwa proses pembelajaran berlangsung secara konsisten dan tidak menunjukkan indikasi *overfitting* yang kuat. Hasil analisis kurva pelatihan tersebut disajikan pada Gambar 5.



Gambar 5 Kurva pelatihan *XGBoost* pada *multilingual e5 large instruct* dengan *borderline-smote*

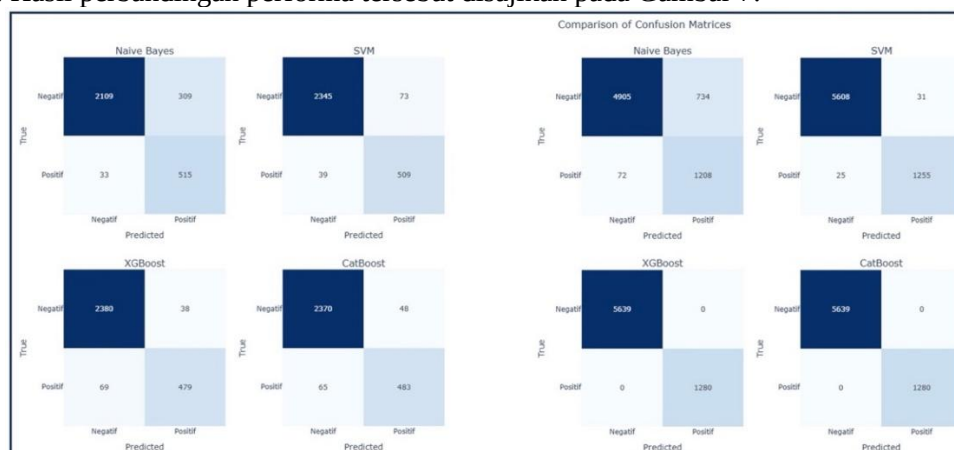
Kurva pelatihan model *CatBoost* dengan representasi *Multilingual E5 Large Instruct* digunakan untuk menganalisis perkembangan kinerja model selama proses pelatihan setelah penerapan *Borderline-SMOTE* pada data latih. Performa validasi yang relatif stabil terhadap performa pelatihan menunjukkan bahwa proses pembelajaran berlangsung secara konsisten dan tidak menunjukkan indikasi *overfitting* yang kuat. Hasil analisis kurva pelatihan tersebut disajikan pada Gambar 6.



Gambar 6 Kurva pelatihan *catboost* pada *multilingual E5 large instruct* dengan *borderline-SMOTE*

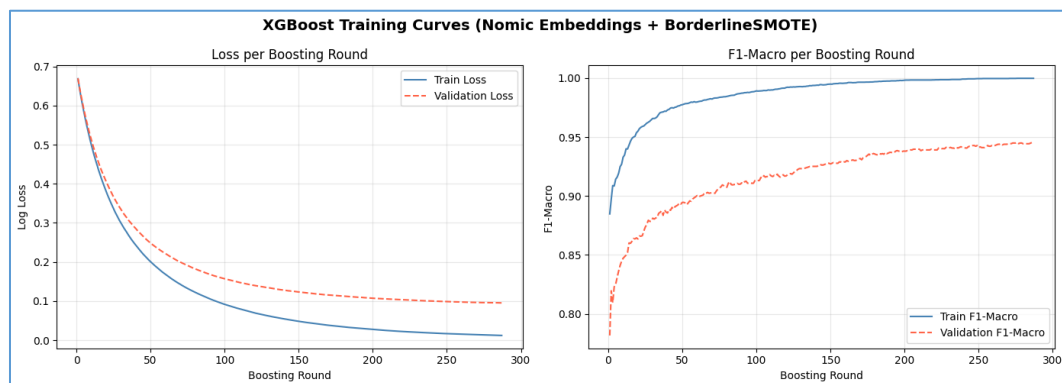
c. Eksperimen Kelompok *Nomic Embed v2*

Perbandingan performa model pada kelompok *Nomic Embed v2* antara data latih dan data uji dilakukan untuk mengevaluasi konsistensi kinerja model pada representasi *embedding Nomic Embed v2*. Selisih performa yang relatif kecil antara data latih dan data uji menunjukkan bahwa model mampu mempertahankan kinerja ketika diterapkan pada data yang tidak digunakan selama proses pelatihan. Hasil perbandingan performa tersebut disajikan pada Gambar 7.



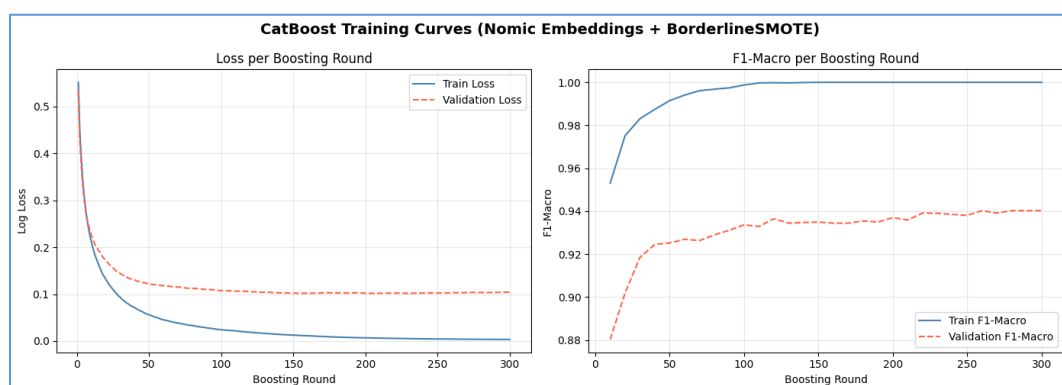
Gambar 7 Hasil perbandingan kelompok *nomic embed v2* pada data uji dan data latih

Kurva pelatihan model *XGBoost* dengan representasi *Nomic Embed v2* digunakan untuk menganalisis perkembangan kinerja model selama proses pelatihan setelah penerapan *Borderline-SMOTE* pada data latih. Performa validasi yang relatif stabil terhadap performa pelatihan menunjukkan bahwa proses pembelajaran berlangsung secara konsisten dan tidak menunjukkan indikasi *overfitting* yang kuat. Hasil analisis kurva pelatihan tersebut disajikan pada Gambar 8.



Gambar 8 Kurva pelatihan *XGBoost* pada *nomic embed v2* dengan *borderline-SMOTE*

Kurva pelatihan model *CatBoost* dengan representasi *Nomic Embed v2* digunakan untuk menganalisis perkembangan kinerja model selama proses pelatihan setelah penerapan *Borderline-SMOTE* pada data latih. Performa validasi yang relatif stabil terhadap performa pelatihan menunjukkan bahwa proses pembelajaran berlangsung secara konsisten dan tidak menunjukkan indikasi *overfitting* yang kuat. Hasil analisis kurva pelatihan tersebut disajikan pada Gambar 9.



Gambar 9. Kurva Pelatihan *CatBoost* pada *Nomic Embed v2* dengan *Borderline-SMOTE*

Secara keseluruhan, Gambar 1 sampai Gambar 9 menunjukkan bahwa performa model pada data latih dan data uji relatif konsisten pada ketiga kelompok representasi teks. Kurva pelatihan *XGBoost* dan *CatBoost* juga memperlihatkan performa validasi yang relatif stabil selama proses pelatihan tanpa penurunan yang tajam. Kondisi tersebut menunjukkan bahwa proses pembelajaran model berlangsung secara konsisten dan tidak menunjukkan indikasi *overfitting* yang kuat. Dengan demikian, hasil pada data uji menunjukkan bahwa model mampu mempertahankan performa yang relatif konsisten ketika diterapkan pada data yang tidak digunakan selama proses pelatihan.

4 Kesimpulan

Penelitian ini berhasil melakukan evaluasi terhadap kombinasi *embedding modern* dan algoritma klasifikasi untuk klasifikasi sentimen ulasan aplikasi Livin' by Mandiri. Hasil penelitian menunjukkan bahwa *Multilingual E5 Large Instruct* menghasilkan rata-rata *F1-macro* tertinggi sebesar 0,950 dibandingkan *TF-IDF* dan *Nomic Embed v2* yang masing-masing memperoleh rata-rata sebesar 0,915. Berdasarkan rata-rata *F1-macro* antar algoritma klasifikasi, *SVM* memperoleh nilai tertinggi sebesar 0,943, sedangkan *XGBoost* dan *CatBoost* masing-masing memperoleh 0,937. Meskipun demikian, kombinasi *Multilingual E5 Large Instruct* dan *CatBoost* menghasilkan performa individual terbaik dengan *accuracy* sebesar 98% dan *F1-macro* sebesar 0,97. Hasil tersebut menunjukkan bahwa

<http://sistemasi.ftik.unisi.ac.id>

performa klasifikasi dipengaruhi oleh interaksi antara representasi teks dan algoritma klasifikasi, sehingga pemilihan kombinasi model perlu mempertimbangkan karakteristik data yang digunakan. Penerapan *Borderline-SMOTE* pada data latih digunakan untuk menangani ketidakseimbangan kelas, sedangkan evaluasi generalisasi menunjukkan bahwa model mampu mempertahankan kinerja pada data uji tanpa indikasi *overfitting* yang kuat.

Referensi

- [1] B. Setiawan, "A Review of Sentiment Analysis Applications in Indonesia Between 2023-2024," 2024. DOI: 10.26740/jieet.v8n2.p71-83.
- [2] L. Wang, N. Yang, X. Huang, L. Yang, R. Majumder, and F. Wei, "Multilingual E5 Text Embeddings: A Technical Report," Feb. 2024, [Online]. Available: <http://arxiv.org/abs/2402.05672>
- [3] N. Muennighoff, N. Tazi, L. Magne, and N. Reimers, "MTEB: Massive Text Embedding Benchmark," Mar. 2023, [Online]. Available: <http://arxiv.org/abs/2210.07316>
- [4] H. Han, W.-Y. Wang, and B.-H. Mao, "Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning," 2005. DOI: 10.1007/11538059_91.
- [5] R. Wahyudi et al., "Analisis Sentimen pada Review Aplikasi Grab di Google Play Store Menggunakan Support Vector Machine," *Jurnal Informatika*, Vol. 8, No. 2, 2021, [Online]. Available: <http://ejournal.bsi.ac.id/ejurnal/index.php/ji>
- [6] C. C. Yolanda, S. Syafriandi, Y. Kurniawati, and D. Fitria, "Sentiment Analysis of DANA Application Reviews on Google Play Store Using Naïve Bayes Classifier Algorithm based on Information Gain," *UNP Journal of Statistics and Data Science*, Vol. 2, No. 1, pp. 48–55, Feb. 2024, DOI: 10.24036/ujsds/vol2-iss1/147.
- [7] W. Chen, K. Yang, Z. Yu, Y. Shi, and C. L. P. Chen, "A Survey on Imbalanced Learning: Latest Research, Applications and Future Directions," *Artif. Intell. Rev.*, Vol. 57, No. 6, Jun. 2024, DOI: 10.1007/s10462-024-10759-6.
- [8] M. C. Hinojosa Lee, J. Braet, and J. Springael, "Performance Metrics for Multilabel Emotion Classification: Comparing Micro, Macro, and Weighted F1-Scores," *Applied Sciences (Switzerland)*, Vol. 14, No. 21, Nov. 2024, DOI: 10.3390/app14219863.
- [9] A. Suharman and M. Kamayani Sulaeman, "Analisis Sentimen Pengguna Aplikasi Livin' by Mandiri menggunakan Metode Support Vector Machine (SVM) dengan Ekstraksi Fitur TF-IDF dan Word2Vec," *Jurnal Pendidikan dan Teknologi Indonesia*, Vol. 5, No. 8, pp. 2201–2212, Aug. 2025, DOI: 10.52436/1.jpti.941.
- [10] Q. H. Nguyen et al., "Influence of Data Splitting on Performance of Machine Learning Models in Prediction of Shear Strength of Soil," *Math. Probl. Eng.*, Vol. 2021, 2021, DOI: 10.1155/2021/4832864.
- [11] I. Nyoman Saputra Wahyu Wijaya, K. Agus Seputra, and N. Putu Novita Puspa Dewi, "Fine Tuning Model Indobert untuk Analisis Sentimen Berita Pariwisata Indonesia," *Jurnal Pendidikan Teknologi dan Kejuruan*, Vol. 22, No. 2, 2025, DOI: 10.23887/jptk-undiksha.v22i2.104056.
- [12] Z. Nussbaum and B. Duderstadt, "Training Sparse Mixture Of Experts Text Embedding Models," Mar. 2025, [Online]. Available: <http://arxiv.org/abs/2502.07972>
- [13] I. B. N. W. Manuaba, Gede Rasben Dantes, and Gede Indrawan, "Analisis Sentimen Data Provider Layanan Internet pada Twitter menggunakan Support Vector Machine (SVM) dengan Penambahan Algoritma Levenshtein Distance," 2022.
- [14] A. Sasmita, G. A. Pradnyana, and D. G. H. Divayana, "Pengembangan Sistem Analisis Sentimen untuk Evaluasi Kinerja Dosen Universitas Pendidikan Ganesha dengan Metode Naïve Bayes," *JST (Jurnal Sains dan Teknologi)*, Vol. 11, No. 2, pp. 451–462, Sep. 2022, DOI: 10.23887/jstundiksha.v11i2.44384.
- [15] N. K. T. A. Saputri, I. G. A. Gunadi, and I. M. G. Sunarya, "Analisis Sentimen Pelayanan Daring di Fakultas Teknik dan Kejuruan Universitas Pendidikan Ganesha menggunakan Algoritma Naïve Bayes dan LSTM," *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, Vol. 4, No. 3, pp. 1120–1129, Jul. 2024, DOI: 10.57152/malcom.v4i3.1336.

- [16] P. W. Ariyani, I. Made, G. Sunarya, I. Gede, and A. Gunadi, “Analisis Sentimen Masyarakat terhadap Virus Corona berdasarkan Opini dari Twitter menggunakan Metode Naive Bayes dan K-Nearest Neighbor,” *Jurnal Pendidikan Teknologi dan Kejuruan*, Vol. 22, No. 2, 2025.
- [17] N. H. Cahyana, Y. Fauziah, W. Wisnalmawati, A. S. Aribowo, and S. Saifullah, “The Evaluation of Effects of Oversampling and Word Embedding on Sentiment Analysis,” *JURNAL INFOTEL*, Vol. 17, No. 1, pp. 54–67, Apr. 2025, DOI: 10.20895/infotel.v17i1.1077.
- [18] J. Snoek, H. Larochelle, and R. P. Adams, “Practical Bayesian Optimization of Machine Learning Algorithms,” 2012, pp. 2951–2959.
- [19] E. Lopez, J. Etxebarria-Elezgarai, J. M. Amigo, and A. Seifert, “The Importance of Choosing a Proper Validation Strategy in Predictive Models. A Tutorial with Real Examples,” Sep. 22, 2023, Elsevier B.V. DOI: 10.1016/j.aca.2023.341532.
- [20] T. Fearn, “Testing Differences in Predictive Ability: A Tutorial,” *J. Chemom.*, Vol. 38, No. 8, Aug. 2024, DOI: 10.1002/cem.3549.
- [21] L. A. Yates, Z. Aandahl, S. A. Richards, and B. W. Brook, “Cross Validation for Model Selection: A Review with Examples from Ecology,” *Ecol. Monogr.*, Vol. 93, No. 1, Feb. 2023, DOI: 10.1002/ecm.1557.