e-*ISSN 2548-6438*
p-ISSN 2614-7173

EVALUASI MODEL MACHINE LEARNING, DEEP LEARNING, DAN TRANSFORMER UNTUK ANALISIS SENTIMEN ULASAN SPKLU PADA PLN MOBILE DENGAN PENANGANAN KETIDAKSEIMBANGAN KELAS

EVALUATION OF MACHINE LEARNING, DEEP LEARNING, AND TRANSFORMER MODELS FOR SENTIMENT ANALYSIS OF SPKLU REVIEWS ON THE PLN MOBILE APP WITH CLASS IMBALANCE HANDLING

Haidi Fildzah Izzati [1], Riska Dhenabayu [2]
Fakultas Ekonomika dan Bisnis, Program Studi Bisnis Digital
Universitas Negeri Surabaya
Email: haidi.23124@mhs.unesa.ac.id

Abstrak

Fitur *EV Charging/SPKLU* pada aplikasi PLN Mobile mendukung layanan pengisian daya kendaraan listrik, tetapi banyaknya ulasan pengguna menyulitkan identifikasi sentimen secara manual. Penelitian ini bertujuan menganalisis sentimen ulasan pengguna terkait fitur *EV Charging/SPKLU* dan membandingkan kinerja tujuh algoritma klasifikasi. Data diperoleh dari *Google Play Store*, kemudian melalui tahap *text preprocessing* dan pelabelan sentimen berdasarkan rating sebagai *weak label*. Algoritma yang digunakan meliputi *Logistic Regression, Support Vector Machine (SVM), Naive Bayes, Multi-Layer Perceptron (MLP), Recurrent Neural Network (RNN), Bidirectional Long Short-Term Memory (BiLSTM)*, dan *DistilBERT*, yang dikelompokkan menjadi *Machine Learning, Deep Learning*, dan *Transformer*. Untuk mengatasi ketidakseimbangan kelas, diterapkan pembobotan kelas atau *class weighting* yang disesuaikan dengan karakteristik masing-masing algoritma pada data latih. Evaluasi dilakukan menggunakan *Accuracy, Precision, Recall, Macro F1-Score*, dan *confusion matrix*. Hasil menunjukkan bahwa MLP memperoleh *Macro F1-Score* tertinggi pada data validasi sebesar 67,80% sehingga dipilih untuk pengujian akhir. Pada 286 data uji, MLP menghasilkan *Accuracy* sebesar 89,16% dan *Macro F1-Score* sebesar 59,42%, yang menunjukkan kinerja klasifikasi keseluruhan yang tinggi dengan performa yang belum merata pada setiap kelas sentimen.

Kata Kunci: Analisis Sentimen, SPKLU, PLN Mobile, *Imbalanced Learning, DistilBERT*

Abstract

The EV Charging/SPKLU feature on the PLN Mobile app supports electric vehicle charging services, but the large number of user reviews makes it difficult to identify sentiment manually. This study aims to analyze the sentiment in user reviews related to the EV Charging/SPKLU feature and compare the performance of seven classification algorithms. Data was obtained from the Google Play Store, then underwent text preprocessing and sentiment labeling based on ratings as weak labels. The algorithms used include Logistic Regression, Support Vector Machine (SVM), Naive Bayes, Multi-Layer Perceptron (MLP), Recurrent Neural Network (RNN), Bidirectional Long Short-Term Memory (BiLSTM), and DistilBERT, which are categorized into Machine Learning, Deep Learning, and Transformer. To address class imbalance, class weighting was applied, adjusted according to the characteristics of each algorithm on the training data. The evaluation was conducted using Accuracy, Precision, Recall, Macro F1-Score, and a confusion matrix. The results show that the MLP achieved the highest Macro F1-Score on the validation data at 67.80%, so it was selected for the final testing. On the 286 test data points, the MLP achieved an accuracy of 89.16% and a Macro F1-Score of 59.42%, indicating high overall classification performance, though performance was not consistent across all sentiment classes.

Keywords: Sentiment Analysis, SPKLU, PLN Mobile, *Imbalanced Learning, DistilBERT*

PENDAHULUAN

Kendaraan listrik (*Electric Vehicle*/EV) semakin berkembang di Indonesia seiring dengan pembangunan infrastruktur Stasiun Pengisian Kendaraan Listrik Umum (SPKLU). PT PLN (Persero) turut mendukung perkembangan ekosistem kendaraan listrik melalui layanan EV Charging/SPKLU pada aplikasi PLN Mobile yang memungkinkan pengguna memperoleh informasi dan memanfaatkan layanan pengisian kendaraan listrik secara digital. Sebagai aplikasi yang memungkinkan pengguna memberikan ulasan dan *rating* melalui Google Play Store, PLN Mobile menghasilkan data opini yang dapat digunakan untuk memahami pengalaman pengguna terhadap layanan tersebut. Ulasan pengguna dapat memuat tanggapan mengenai kemudahan penggunaan, ketersediaan SPKLU, maupun kendala yang dialami selama menggunakan layanan EV Charging/SPKLU. Dengan jumlah ulasan yang terus bertambah, analisis sentimen diperlukan untuk

mengidentifikasi kecenderungan opini pengguna secara sistematis.

Banyaknya ulasan pengguna menyebabkan proses identifikasi dan pengelompokan opini secara manual menjadi kurang efisien. Analisis sentimen dapat digunakan untuk mengklasifikasikan ulasan secara otomatis ke dalam kelas *Positive, Neutral,* dan *Negative*, sehingga opini pengguna dapat dianalisis dalam jumlah data yang besar. Namun, data ulasan pada penelitian ini memiliki distribusi kelas yang tidak seimbang, dengan jumlah ulasan *Positive* jauh lebih banyak dibandingkan kelas *Negative* dan *Neutral*. Kondisi tersebut dapat menyebabkan model lebih dominan mempelajari kelas mayoritas sehingga kemampuan dalam mengenali kelas dengan jumlah data lebih sedikit menjadi kurang optimal. Oleh karena itu, karakteristik distribusi kelas menjadi salah satu aspek penting yang perlu diperhatikan dalam proses klasifikasi ulasan EV Charging/SPKLU.

Ketidakseimbangan kelas memerlukan strategi penanganan yang disesuaikan dengan karakteristik algoritma yang digunakan. Pada penelitian ini, penanganan *class imbalance* diterapkan pada data latih melalui pembobotan yang disesuaikan dengan masing-masing algoritma. Logistic Regression dan SVM menggunakan *class weight*, Naive Bayes menggunakan *class prior* yang dibuat seragam, RNN, BiLSTM, dan MLP menggunakan *class weight*, sedangkan DistilBERT menggunakan *weighted cross-entropy loss*. Data validasi dan data uji tidak mengalami perubahan distribusi sehingga tetap merepresentasikan kondisi data sebenarnya. Selain itu, *Macro F1-Score* digunakan sebagai dasar pemilihan model karena memberikan bobot yang sama pada setiap kelas, sehingga evaluasi tidak hanya dipengaruhi oleh kelas mayoritas.

Penelitian mengenai analisis sentimen pada aplikasi PLN Mobile telah dilakukan menggunakan berbagai algoritma klasifikasi yang menerapkan VADER dan Naive Bayes pada 1.000 ulasan PLN Mobile dengan tiga kelas sentimen, yaitu *Positive, Neutral,* dan *Negative*.(Asri et al., 2022). Penelitian lain menggunakan Naive Bayes Classifier dan K-Nearest Neighbor terhadap 3.000 ulasan PLN Mobile dan memperoleh distribusi 69,97% *Positive*, 12,27% *Neutral*, dan 17,77% *Negative*, dengan akurasi Naive Bayes sebesar 77,69% (Afdal & Novita, 2024). Penelitian yang lebih baru membandingkan Naive Bayes, SVM, dan Random Forest pada ulasan PLN Mobile serta menerapkan SMOTE untuk menangani ketidakseimbangan kelas. Penelitian tersebut menggunakan dua kelas sentimen dan

memperoleh akurasi terbaik sebesar 84% pada pengujian.(Agta et al., 2026). Penelitian lainnya juga membandingkan Naive Bayes, Random Forest, dan SVM pada 19.870 ulasan PLN Mobile serta menunjukkan peningkatan kinerja setelah penerapan SMOTE. (Nurita et al., 2025).

Meskipun berbagai penelitian telah membahas analisis sentimen pada ulasan PLN Mobile, penelitian terdahulu umumnya menggunakan ulasan aplikasi secara umum dan belum secara khusus memusatkan analisis pada fitur EV Charging/SPKLU. Selain itu, penelitian yang telah menerapkan penanganan *class imbalance* pada PLN Mobile masih menggunakan jumlah algoritma dan konfigurasi evaluasi yang lebih terbatas, seperti penggunaan SMOTE pada model *Machine Learning* dan klasifikasi dua kelas sentimen (Agta et al., 2026). Dengan demikian, masih diperlukan evaluasi yang secara khusus menguji berbagai pendekatan *Machine Learning, Deep Learning,* dan *Transformer* pada ulasan EV Charging/SPKLU PLN Mobile dengan tiga kelas sentimen yang distribusinya tidak seimbang. Perbedaan karakteristik objek, distribusi kelas, penanganan *imbalance*, serta strategi evaluasi tersebut menjadi dasar dilakukannya penelitian ini.

Berdasarkan kesenjangan tersebut, penelitian ini melakukan evaluasi komparatif tujuh model pada domain spesifik ulasan EV Charging/SPKLU PLN Mobile dengan distribusi kelas yang tidak seimbang, yaitu *Logistic Regression, Support Vector Machine (SVM), Naive Bayes, Multi-Layer Perceptron (MLP), Recurrent Neural Network (RNN), Bidirectional Long Short-Term Memory (BiLSTM),* dan *DistilBERT*. Penanganan *class imbalance* diterapkan pada data latih sesuai karakteristik masing-masing model, sedangkan pemilihan model dilakukan berdasarkan *Macro F1-Score* pada data validasi. Model dengan kinerja terbaik pada data validasi kemudian dilakukan pengujian akhir menggunakan data uji yang benar-benar terpisah dari proses pelatihan dan pemilihan model. Dengan rancangan tersebut, penelitian ini memberikan bukti empiris mengenai kinerja relatif tujuh model dalam klasifikasi sentimen ulasan EV Charging/SPKLU PLN Mobile pada kondisi distribusi kelas yang tidak seimbang.

LANDASAN TEORI

Analisis Sentimen

Analisis Sentimen merupakan bidang dalam *text mining* yang

memanfaatkan pendekatan komputasi untuk mengidentifikasi dan menganalisis opini, sentimen, serta subjektivitas yang terdapat dalam teks terhadap suatu objek atau entitas(Medhat et al., 2014)

Preprocessing Text

merupakan tahapan awal untuk mempersiapkan teks agar dapat diproses dalam analisis. Tahapan ini meliputi *cleaning* untuk menghapus unsur yang tidak diperlukan, *tokenizing* untuk memecah teks menjadi kata, *stopword removal* untuk menghilangkan kata yang kurang informatif, dan *stemming* untuk mengubah kata berimbuhan menjadi bentuk dasarnya.

Machine Learning

Machine Learning merupakan bidang yang mempelajari kemampuan komputer agar dapat "belajar" yaitu merubah data menjadi suatu keahlian berupa program/model yang dapat menjalankan tugas tertentu secara otomatis, terutama pada permasalahan yang sulit diselesaikan melalui pemrograman secara eksplisit oleh manusia (Shalev-Shwartz & Ben-David, 2014).

Dalam analisis sentimen, teks perlu direpresentasikan dalam bentuk numerik sebelum diproses oleh algoritma *Machine Learning*. Salah satu metode yang digunakan adalah *Term Frequency-Inverse Document Frequency* (TF-IDF), yang memberikan bobot pada kata berdasarkan kemunculannya dalam dokumen dan keseluruhan kumpulan dokumen. Hasil pembobotan TF-IDF selanjutnya dapat digunakan sebagai fitur masukan bagi algoritma klasifikasi seperti *Naïve Bayes*(Wati et al., 2023)

Deep Learning

Deep Learning merupakan bagian pendekatan Machine Learning yang menggunakan beberapa lapisan pemrosesan untuk mempelajari representasi data secara bertingkat, mulai dari representasi yang lebih sederhana sampai representasi yang lebih kompleks (LeCun & Hinton, 2015)

Transformer

Transformer merupakan arsitektur jaringan saraf yang diperkenalkan oleh (Vaswani, 2017), yang memanfaatkan mekanisme *self-attention* untuk memahami hubungan antar kata dalam teks. Arsitektur ini kemudian menjadi dasar bagi berbagai model bahasa, termasuk BERT dan *DistilBERT. DistilBERT* yaitu model berbasis *Transformer* yang dikembangkan dari BERT menggunakan teknik *knowledge distillation* sehingga memiliki ukuran lebih kecil dan proses inferensi lebih

cepat, namun tetap mempertahankan kemampuan pemahaman bahasa (Sanh et al., 2019). Dalam penelitian ini, DistilBERT digunakan sebagai salah satu algoritma pada kelompok Transformer untuk klasifikasi sentimen ulasan PLN Mobile.

Ketidakseimbangan Kelas *(Class Imbalance)*

Ketidakseimbangan kelas adalah masalah yang sering mundul dalam Machine Learning yang muncul dengan rasio instance yang tidak proposional. (Bi & Zhang, 2018)

Kondisi ini dapat menyebabkan model lebih cenderung mengenali kelas mayoritas. Penanganannya dapat dilakukan melalui *class weighting* dan *cost-sensitive learning* dengan memberikan bobot atau penalti lebih besar pada kelas minoritas, serta *weighted loss* pada fungsi kerugian. Dalam penelitian ini, *Logistic Regression, SVM, RNN, BiLSTM,* dan *MLP* menggunakan pembobotan kelas, *Naive Bayes* menggunakan *class prior* seragam, sedangkan *DistilBERT* menggunakan *weighted cross-entropy loss*.

Macro F1-Score

Macro F1-Score merupakan metrik evaluasi pada klasifikasi multikelas yang diperoleh dengan menghitung nilai F1 untuk setiap kelas, kemudian mengambil rata-rata dari seluruh nilai tersebut dengan bobot yang sama (Takahashi et al., 2022). Dengan demikian, kinerja setiap kelas tetap diperhitungkan tanpa bergantung pada jumlah data pada masing-masing kelas. Karakteristik tersebut membuat *Macro F1-Score* sesuai digunakan pada dataset dengan distribusi kelas tidak seimbang, karena performa pada kelas mayoritas dan minoritas tetap memiliki kontribusi yang sama dalam hasil evaluasi.

METODE PENELITIAN

Penelitian ini menggunakan pendekatan kuantitatif eksperimental untuk membandingkan kinerja beberapa algoritma klasifikasi dalam menganalisis sentimen ulasan pengguna aplikasi PLN Mobile yang berkaitan dengan layanan Kendaraan listrik Electric Vehicle (EV) Charging/Stasiun Pengisian Kendaraan Listrik Umum (SPKLU).

Perbandingan dilakukan berdasarkan metrik evaluasi yang telah terukur, yaitu Accuracy, Precision, Recall, dan F1-Score. Tahapan penelitian meliputi pengumpulan dan seleksi data, *text preprocessing*, pelabelan dan validasi data, pembagian data, representasi fitur, pelatihan model, serta evaluasi.

Pengumpulan dataset

Pengumpulan data dilakukan melalui *web scraping* menggunakan *library* google-play-scraper pada Google Colaboratory. Data berupa ulasan aplikasi PLN Mobile dengan *package ID* com.icon.pln123 yang diambil pada periode 24 Juli 2022–30 Juni 2026. Selanjutnya, ulasan disaring berdasarkan kata kunci terkait EV Charging/SPKLU, seperti *spklu, ev, charging, charger, kendaraan listrik,* dan *pengisian daya* menggunakan *regular expression* berbasis *word boundary*. Ulasan yang memenuhi kriteria digunakan sebagai dataset untuk tahap *preprocessing*.

Preprocessing Text

Dataset hasil penyaringan akan dilanjutkan menuju tahap preprocessing text yang terdiri dari cleaning, tokenizing, stopword removal, stemming.

Tahap *cleaning* dilakukan untuk menghilangkan unsur yang tidak diperlukan dalam teks, seperti URL, simbol, dan tanda baca. *Tokenizing* dilakukan dengan memecah teks menjadi token atau kata. Tahap *stopword removal* digunakan untuk menghilangkan kata-kata yang kurang memberikan informasi dalam proses klasifikasi, sedangkan *stemming* dilakukan untuk mengubah kata berimbuhan menjadi bentuk dasarnya menggunakan pustaka PySastrawi.

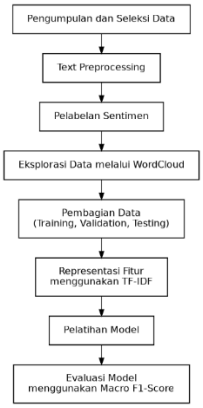
Pelabelan

Pelabelan sentimen dilakukan berdasarkan rating yang diberikan pengguna, dengan rating 1–2 dikategorikan sebagai *Negative*, rating 3 sebagai *Neutral*, dan rating 4–5 sebagai *Positive*. Pendekatan *rating-based labeling* ini termasuk dalam kategori *distant supervision* atau *weak supervision*, yaitu strategi pelabelan yang memanfaatkan sinyal tidak langsung pada data sebagai pengganti anotasi manual terhadap seluruh dataset (Lalor et al., 2023), dan telah diterapkan pada penelitian analisis sentimen ulasan produk marketplace Indonesia menggunakan Naive Bayes dan SVM (Gulo & Wibowo, 2026)

Pelabelan berbasis rating ini memiliki keterbatasan, karena rating yang diberikan pengguna tidak selalu konsisten dengan keseluruhan sentimen yang terkandung dalam teks ulasan (Lalor et al., 2023). Hal ini dapat terjadi karena pengguna tidak menuliskan seluruh aspek pengalamannya dalam ulasan, atau karena sentimen yang diberikan didominasi oleh aspek tertentu saja. Dalam konteks penelitian ini, pengguna dapat memberikan rating rendah karena keluhan terhadap aspek di luar fitur EV Charging/SPKLU itu sendiri, atau memberikan rating tinggi disertai komentar yang mengandung kritik minor.

Analisys Data/Sistem

Sistem klasifikasi sentimen pada penelitian ini dilakukan melalui beberapa tahapan, mulai dari pengumpulan data, pengolahan data, hingga evaluasi model, seperti ditunjukkan pada Gambar 1.



Gambar 1. Tahapan Penelitian

Setelah data melewati tahap *text preprocessing* dan pelabelan sentimen, dataset kemudian dibagi menjadi data latih (*training*), validasi (*validation*), dan uji (*testing*). Pembagian data menggunakan metode *stratified sampling* dengan parameter stratify=y untuk menjaga proporsi kelas *Negative, Neutral,* dan *Positive* pada masing-masing subset. Secara umum, pembagian sampel berdasarkan strata dapat dinyatakan dengan persamaan berikut [Cochran, 1977]:

$$n_h = \frac{N_h}{N} \times n$$

dengan n_h merupakan jumlah sampel pada strata atau kelas ke-h N_h, merupakan jumlah data pada kelas ke-h, N merupakan jumlah keseluruhan data, dan n merupakan jumlah sampel yang dialokasikan.

Pembagian dataset dilakukan dalam dua tahap. Pertama, sebanyak 20% dari keseluruhan data digunakan sebagai data uji, sedangkan 80% sisanya digunakan sebagai data latih dan validasi. Perhitungannya ditunjukkan pada persamaan berikut:

$$N_{test} = 0{,}20N$$
$$N_{train+val} = (1 - 0{,}20)N = 0{,}80N$$

Selanjutnya, data sebesar 80% tersebut dibagi kembali dengan perbandingan 85:15 untuk memperoleh data latih dan validasi. Dengan demikian, jumlah data pada masing-masing subset adalah:

$$N_{train} = 0{,}85 \times 0{,}80N = 0{,}68N$$
$$N_{val} = 0{,}15 \times 0{,}80N = 0{,}12N$$

Dengan demikian, pembagian akhir dataset dapat dinyatakan sebagai:

68% Training+12% Validation+20% Testing=100%

Nilai random_state=42 digunakan untuk mengontrol proses pengacakan sehingga pembagian dataset dapat direproduksi pada eksperimen berikutnya. Data latih digunakan untuk proses pembelajaran model, data validasi digunakan untuk membandingkan performa ketujuh model dan menentukan model terbaik, sedangkan data uji disimpan hingga tahap akhir dan digunakan satu kali untuk memperoleh performa akhir model terpilih.

Evaluasi dan Pemilihan Model

Model terbaik ditentukan berdasarkan Macro F1-Score pada data validation, bukan berdasarkan data testing. Pemilihan Macro F1 digunakan karena penelitian memiliki distribusi kelas yang tidak seimbang sehingga evaluasi tidak hanya berfokus pada kelas mayoritas.

Setelah model terbaik diperoleh, model tersebut diuji menggunakan data testing yang sebelumnya tidak digunakan dalam proses pelatihan maupun pemilihan model. Hasil pada data testing menjadi performa akhir model dalam penelitian.

HASIL DAN PEMBAHASAN

Pengumpulan Data

Pengumpulan data dilakukan melalui proses *scraping* otomatis menggunakan *library* google-play-scraper pada platform Google Colaboratory. Data yang dikumpulkan berupa ulasan pengguna aplikasi PLN Mobile dengan identifikasi paket com.icon.pln123 dalam rentang waktu 24 Juli 2022 hingga 30 Juni 2026. Pengambilan data menggunakan parameter bahasa Indonesia (lang='id') dan wilayah Indonesia (country='id'), serta diurutkan berdasarkan ulasan terbaru menggunakan Sort.NEWEST. Berdasarkan rentang waktu tersebut, diperoleh sebanyak 541.373 ulasan. Selanjutnya, dilakukan penyaringan untuk memperoleh ulasan yang berkaitan dengan fitur EV Charging/SPKLU menggunakan kata kunci *spklu, ev, charging, charger, charge, cas, kendaraan listrik, mobil listrik, motor listrik, stasiun pengisian, electric vehicle, pengisian listrik,* dan *pengisian daya*. Hasil penyaringan menghasilkan 1.426 ulasan yang relevan dengan topik EV Charging/SPKLU. Distribusi rating pada data hasil penyaringan terdiri atas :

Tabel 1. Hasil Scraping Dataset

Tahapan	Jumlah
Dataset Seluruh Ulasan PLN Mobile	541.373
Dataset relevan EV Charging/SPKLU	1.426

Hasil tersebut menunjukkan bahwa hanya sebagian kecil ulasan PLN Mobile yang secara eksplisit berkaitan dengan EV Charging/SPKLU. Oleh karena itu, filtering diperlukan agar model mempelajari opini yang sesuai dengan ruang lingkup penelitian dan tidak tercampur dengan ulasan fitur lain

Hasil Preprocessing Text

Tahap preprocessing dilakukan untuk mengubah ulasan pengguna yang masih berupa teks mentah menjadi teks yang lebih terstruktur dan siap digunakan dalam proses klasifikasi. Tahapan yang diterapkan meliputi *cleaning*, *tokenizing*, *stopword removal*, *stemming*, dan *joining*. Pada tahap *cleaning*, karakter yang tidak diperlukan seperti URL,

angka, emoji, tanda baca, dan simbol dihapus serta teks diubah menjadi huruf kecil. Selanjutnya, teks dipecah menjadi token kata, kemudian kata yang termasuk *stopword* dihapus dengan tetap mempertahankan kata-kata yang memiliki makna sentimen dan kata kunci domain EV/SPKLU.

Tahap *stemming* kemudian digunakan untuk mengubah kata ke bentuk dasarnya menggunakan PySastrawi. Setelah proses tersebut selesai, token digabungkan kembali menjadi teks untuk digunakan sebagai masukan model klasifikasi. Tahap preprocessing tidak mengubah jumlah data secara substantif, tetapi bertujuan mengurangi variasi bentuk teks dan karakter yang tidak relevan sehingga fitur yang digunakan model menjadi lebih terstruktur.

Tabel 2. Hasil Preprocessing Text

Tahap	Hasil
Komentar Asli	aplikasi yang begitu membantu, ditambah dengan ada nya fitur2 yang mumpuni seperti plan trip yg dapat memudahkan, tanpa lagi ragu dalam menggunakan kendaraan EV..karena SPKLU sudah tersedia tersebar..
Cleaning	aplikasi yang begitu membantu ditambah dengan ada nya fitur yang mumpuni seperti plan trip yg dapat memudahkan tanpa lagi ragu dalam menggunakan kendaraan ev karena spklu sudah tersedia tersebar
Tokenizing	['aplikasi', 'yang', 'begitu', 'membantu', 'ditambah', 'dengan', 'ada', 'nya', 'fitur', 'yang', 'mumpuni', 'seperti', 'plan', 'trip', 'yg', 'dapat', 'memudahkan', 'tanpa', 'lagi', 'ragu', 'dalam', 'menggunakan', 'kendaraan', 'ev', 'karena', 'spklu', 'sudah', 'tersedia', 'tersebar']
Stopword Removal	['aplikasi', 'membantu', 'ditambah', 'fitur', 'mumpuni', 'plan', 'trip', 'memudahkan', 'ragu', 'kendaraan', 'ev', 'spklu', 'tersedia', 'tersebar']
Stemming	['aplikasi', 'bantu', 'tambah', 'fitur', 'mumpuni', 'plan', 'trip', 'mudah', 'ragu', 'kendara', 'ev', 'spklu', 'sedia', 'sebar']
Join	aplikasi bantu tambah fitur mumpuni plan trip mudah ragu kendara ev spklu sedia sebar

Hasil Pelabelan

Pelabelan dilakukan secara otomatis berdasarkan rating pengguna. Rating 1–2 dikategorikan sebagai Negative, rating 3 sebagai Neutral, dan rating 4–5 sebagai Positive. Label tersebut merupakan pseudo-label berbasis rating.

Tabel 3. Label Sentimen

Rating	Jumlah	Label Sentimen	Persentase
1-2	189	Negative	13,25%
3	44	Neutral	3,09%
4-5	1.193	Positive	83,66%
Total	1.426		100%

Persentase kelas dihitung dengan $\text{Persentase}\_i = (N_i/N) \times 100$ dengan N_i adalah jumlah data kelas ke-i dan N adalah total data. Berdasarkan hasil tersebut, kelas Positive mendominasi dengan 1.193 ulasan atau 83,66%, sedangkan Negative berjumlah 189 ulasan atau 13,25% dan Neutral hanya 44 ulasan atau 3,09%. Distribusi ini menunjukkan class imbalance yang kuat.

Ketidakseimbangan kelas penting dalam interpretasi performa

model. Model dapat menghasilkan accuracy tinggi karena dominasi kelas mayoritas, sementara kemampuan mengenali kelas minoritas belum tentu baik. Oleh karena itu, hasil penelitian perlu dibaca bersama Precision, Recall, F1-Score, dan terutama Macro F1.

Data Training, Validation, dan Testing

Pembagian data pada penelitian ini dilakukan melalui dua tahap splitting secara berurutan, dengan tujuan memastikan data uji benar-benar terpisah dan tidak digunakan dalam proses pelatihan maupun validasi model.

a.) Tahap awal, seluruh data dibagi menjadi dua bagian dengan perbandingan 80:20, yaitu 80% data latih sementara (temporary training set) dan 20% data uji (testing set). Tahap pertama ini akan memisahkan 20% data sebagai data uji, sehingga sisa data latih-validasi adalah :

$$N_{test} = 0{,}20N$$
$$N_{train+val} = (1 - 0{,}20)N = 0{,}80N$$

b.) Tahap kedua, 80% data latih sementara dari tahap pertama dibagi kembali dengan perbandingan 85:15, yang akan menghasilkan data latih (training set) sebesar 85% dan data validasi sebesar 15% dari data latih sementara.

$$N_{train} = 0{,}85 \times 0{,}80N = 0{,}68N$$
$$N_{val} = 0{,}15 \times 0{,}80N = 0{,}12N$$

Kedua tahap pembagian ini dilakukan secara stratified berdasarkan label sentiment, dengan random state_42, sehingga proporsi tiap kelas tetap terjaga pada ketiga subset dan hasil pembagian dapat direproduksi.

Tabel 4. Data Training, Validation, dan Testing

Subset	Proporsi	Jumlah	Kegunaan
Training	68%	969	Melatih Model
Validation	12%	171	Memilih model terbaik
Testing	20%	286	Evaluasi akhir

Pembobotan TF-IDF

Fitur teks hasil preprocessing diubah menjadi representasi numerik menggunakan *Term Frequency–Inverse Document Frequency* (TF-IDF), yang memberi bobot tinggi pada term yang sering muncul dalam suatu dokumen namun jarang muncul di dokumen lain. Pembobotan dilakukan menggunakan TfidfVectorizer (scikit-learn, max_features=3000), yang di-*fit* hanya pada data latih untuk mencegah *data leakage* ke data validasi dan uji.

$$TF(t,d) = f(t,d)$$

$$IDF(t) = \ln\left(\frac{1+N}{1+df(t)}\right) + 1$$

$$w(t,d) = TF(t,d) \times IDF(t)$$

$$TFIDF(t,d) = \frac{w(t,d)}{\sqrt{\sum_{t'} w\,(t',d)^2}}$$

dengan t = term, d = dokumen, $f(t,d)$= frekuensi term t pada dokumen d, N = jumlah dokumen data

latih, $df(t)$ = jumlah dokumen yang mengandung term t.

Proses ini menghasilkan 2.066 fitur TF-IDF dari data latih. Tabel menampilkan contoh sepuluh term dengan bobot tertinggi pada satu sampel data latih.

Tabel Sepuluh Term dengan Bobot TF-IDF Tertinggi (Sampel Data Latih ke-1)

Tabel 5. Hasil Pembobotan TF-IDF

Kata	Nilai	TF-IDF
0	admnya	0.496657
1	prabayar	0.420707
2	murah	0.378804
3	lumayan	0.362132
4	token	0.242506
5	baik	0.212463
6	isi	0.198641
7	mobile	0.182937
8	bantu	0.181837
9	mudah	0.172333

Hasil dan Analisis Perbandingan Kinerja Model

Dataset yang digunakan memiliki distribusi kelas yang tidak seimbang (*imbalanced*), dengan jumlah ulasan *Positive* jauh lebih banyak dibandingkan *Negative* dan *Neutral*. Kondisi ini dapat membuat model lebih cenderung mengenali kelas mayoritas. Oleh karena itu, *class imbalance* ditangani pada proses pelatihan sesuai algoritma yang digunakan, yaitu *class weight* pada Logistic Regression, SVM, RNN, BiLSTM, dan MLP, *class prior* seragam pada Naive Bayes, serta *weighted cross-entropy loss* pada DistilBERT. Pembobotan dihitung berdasarkan data latih tanpa mengubah distribusi data validasi dan uji.

Karena distribusi kelas tidak seimbang, *Accuracy* belum cukup untuk menilai kemampuan model pada seluruh kelas. Oleh karena itu, *Macro F1-Score* digunakan sebagai metrik utama karena menghitung rata-rata *F1-Score* setiap kelas dengan bobot yang sama. Evaluasi ketujuh algoritma kemudian dilakukan menggunakan data validasi dan hasil perbandingannya ditampilkan pada Tabel 6.

Tabel 6. Komparasi Metrik Pada Validation Set

No.	Model	Akurasi	Precision	Recall	F1-Score
1	MLP	89.47	66.10	74.42	67.80
2	SVM	87.72	61.83	60.86	60.58
3	Logistic Regression	87.72	57.54	62.07	59.31
4	BiLSTM	84.21	61.04	61.03	57.36
5	Naïve Bayes	91.23	56.13	54.61	55.23
6	RNN	88.30	53.57	53.44	53.49
7	DistilBERT	83.04	46.86	56.21	49.63

Berdasarkan Tabel 6, kinerja ketujuh algoritma dibandingkan menggunakan *Accuracy, Precision, Recall*, dan *Macro F1-Score*. *Accuracy* menunjukkan proporsi prediksi yang benar, *Precision* menunjukkan ketepatan prediksi suatu kelas, sedangkan *Recall* menunjukkan kemampuan model mengenali data pada kelas tersebut. Keduanya dirangkum dalam *F1-Score* untuk menggambarkan keseimbangan *Precision* dan *Recall*. Karena

distribusi data tidak seimbang dan didominasi kelas *Positive*, *Macro F1-Score* digunakan sebagai acuan utama karena memberikan bobot yang sama pada setiap kelas. Oleh karena itu, *Accuracy* yang tinggi belum tentu menunjukkan kinerja yang merata pada seluruh kelas.

Perbedaan peringkat *Accuracy* dan *Macro F1-Score* pada Tabel 6 menunjukkan pentingnya mempertimbangkan kinerja setiap kelas. Berdasarkan *Macro F1-Score*, MLP memperoleh nilai tertinggi sehingga dipilih sebagai model terbaik pada data validasi dan selanjutnya digunakan untuk pengujian pada data uji.

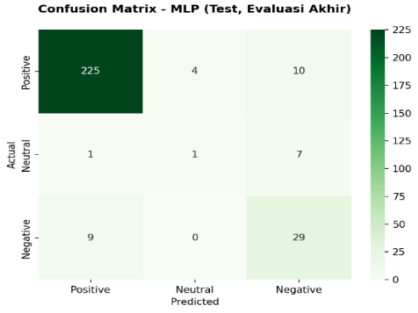
Hasil Pengujian dan Analisis Model Terbaik

Setelah MLP memperoleh *Macro F1-Score* tertinggi pada data validasi, yaitu 67,80%, model tersebut dipilih untuk pengujian akhir menggunakan 286 data uji yang tidak digunakan selama proses pelatihan maupun pemilihan model. Hasil pengujian dibandingkan dengan kinerja pada data validasi untuk melihat kemampuan model dalam mengklasifikasikan data yang belum pernah digunakan sebelumnya.

Tabel 7. Hasil Evaluasi Model

Metrik	Validasi	Data Uji
Accuracy	89,47%	89,16%
Macro F1-Score	67,80%	59,42%

Berdasarkan hasil pengujian, *Accuracy* MLP mengalami penurunan yang relatif kecil, dari 89,47% pada data validasi menjadi 89,16% pada data uji. Sementara itu, *Macro F1-Score* mengalami penurunan yang lebih besar, dari 67,80% menjadi 59,42%. Perbedaan tersebut menunjukkan bahwa kemampuan model dalam menghasilkan prediksi secara keseluruhan masih relatif terjaga, tetapi kinerja dalam mengenali setiap kelas secara seimbang mengalami penurunan yang lebih besar.



Gambar 2. Confusion Matrix

Hasil *confusion matrix* menunjukkan bahwa kelas *Neutral* merupakan kelas yang paling sulit dikenali oleh model. Dari 9 data *Neutral* pada data uji, hanya 1 yang berhasil dikenali dengan benar, sedangkan 1 diprediksi sebagai *Positive* dan 7 sebagai *Negative*. Jumlah data *Neutral* yang hanya 3,09% dari keseluruhan data membuat pola yang tersedia untuk dipelajari model relatif terbatas dibandingkan kelas *Positive* dan *Negative*. Selain itu, penggunaan *pseudo-label*

berdasarkan *rating* memungkinkan adanya ketidaksesuaian antara rating pengguna dan sentimen yang disampaikan dalam teks. Kondisi tersebut dapat menjadi salah satu faktor yang memengaruhi kemampuan model dalam mengenali kelas *Neutral*. Karena *Macro F1-Score* memberikan bobot yang sama pada setiap kelas, rendahnya kinerja pada *Neutral* berdampak lebih besar terhadap nilai *Macro F1-Score*.

KESIMPULAN

Penelitian ini melakukan analisis sentimen pada ulasan pengguna terkait fitur EV Charging/SPKLU pada aplikasi PLN Mobile dengan membandingkan tujuh algoritma dari pendekatan *Machine Learning*, *Deep Learning*, dan *Transformer* serta mempertimbangkan ketidakseimbangan kelas. Hasil evaluasi pada *validation set* menunjukkan bahwa MLP memperoleh *Macro F1-Score* tertinggi sebesar 67,80% dan dipilih untuk pengujian akhir. Pada *test set*, MLP memperoleh *Accuracy* sebesar 89,16% dan *Macro F1-Score* sebesar 59,42%, dengan kelas *Neutral* menjadi kelas yang paling sulit dikenali, yaitu hanya 1 dari 9 data yang berhasil diklasifikasikan dengan benar. Hasil tersebut perlu diinterpretasikan dengan mempertimbangkan penggunaan *pseudo-label* berdasarkan *rating* yang belum sepenuhnya merepresentasikan sentimen pada teks. Penelitian selanjutnya disarankan untuk melakukan validasi label melalui anotasi manusia, membandingkan strategi penanganan *class imbalance* untuk meningkatkan pengenalan kelas *Neutral*, serta menerapkan *Stratified K-Fold Cross-Validation* dan *hyperparameter tuning* untuk memperoleh evaluasi yang lebih stabil.

DAFTAR PUSTAKA

Afdal, M., & Novita, R. (2024). *Sentiment Analysis of PLN Mobile Application Review Using Naïve Bayes Classifier and K-Nearest Neighbor Algorithm Analisis Sentimen Ulasan Aplikasi PLN Mobile Menggunakan Algoritma Naïve Bayes Classifier dan K-Nearest Neighbor*. 4(January), 10–19.

Agta, H., Fatah, A., & Redjeki, R. S. (2026). *Analisis Sentimen Ulasan Aplikasi PLN Mobile Menggunakan Naïve Bayes , SVM , dan Random Forest dengan SMOTE*. 24(1), 32–44.

Asri, Y., Suliyanti, W. N., Kuswardani, D., & Fajri, M. (2022). *Pelabelan Otomatis Lexicon Vader dan Klasifikasi Naive Bayes dalam menganalisis sentimen data ulasan PLN*

Mobile. *15*(2), 264–275.

Bi, J., & Zhang, C. (2018). *An empirical comparison on state-of-the-art multi-class imbalance learning algorithms and a new diversified ensemble learning scheme*. Knowledge-Based Systems, 158, 81–93. *https://doi.org/10.1016/j.knosys.2018.05.037*

Gulo, L. A., & Wibowo, A. (2026). *Analisis Sentimen Ulasan Produk Marketplace Indonesia Menggunakan Naive Bayes dan SVM dengan Label Berdasarkan Rating*. 410–419. *https://doi.org/10.33364/algoritma/v.23-1.3206*

Lalor, J. P., Chen, Y., & Kanuri, V. K. (2023). *I MPLEMENTATION OF U NIFIED S ENTIMENT A NALYSIS WITH*. 1–19.

LeCun, Y., & Hinton, G. (2015). *Deep Learning*. Nature, *521*, 436–444. *https://doi.org/10.1038/nature14539*

Medhat, W., Hassan, A., & Korashy, H. (2014). *Sentiment analysis algorithms and applications: A survey*. Ain Shams Engineering Journal, *5*(4), 1093–1113. *https://doi.org/10.1016/j.asej.2014.04.011*

Nurita, C., Sari, K., & Suryono, R. R. (2025). *Perbandingan Algoritma Naïve Bayes , Random Forest , dan SVM Untuk Analisis Sentiment Aplikasi PLN Mobile Pada Google Play Store*. 7(1). *https://doi.org/10.47065/bits.v7i1.7164*

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). *DistilBERT , a distilled version of BERT : smaller , faster , cheaper and lighter*. 2–6.

Shalev-Shwartz, S., & Ben-David, S. (2014). Introduction. In *Understanding Machine Learning: From Theory to Algorithms* (pp. 1–10). Cambridge University Press.

Takahashi, K., Yamamoto, K., Kuchiba, A., & Koyama, T. (2022). *Confidence interval for micro-averaged F1 and macro-averaged F1 score*s. Applied Intelligence, 52(5), 4961–4972. *https://doi.org/10.1007/s10489-021-02635-5*

Vaswani, A. (2017). *Attention Is All You Need*. (Nips). *https://doi.org/10.48550/arXiv.1706.03762*

Wati, R., Ernawati, S., & Rachmi, H. (2023). *Pembobotan TF-IDF Menggunakan Naïve Bayes pada Sentimen Masyarakat Mengenai Isu Kenaikan BIPIH*. Jurnal Manajemen Informatika (JAMIKA), 13, 84–93. *https://doi.org/10.34010/jamika.v13i1.9424*