

Analisis Sentimen Pengguna terhadap Layanan Mobile Banking di Indonesia Menggunakan Lexicon-Based dan BERT

Siska Mayasari Rambe*, Zulham Sitorus, Muhammad Iqbal

Pascasarjana, Magister Teknologi Informasi, Universitas Pembangunan Panca Budi, Medan, Indonesia

Email: ¹mayasarisiska5@gmail.com, ²zulhamsitorus@dosen.pancabudi.ac.id, ³muhammadiqbal@dosenpancabudi.ac.id

Email Penulis Korespondensi: mayasarisiska5@gmail.com*

Submitted: 29/08/2026; Accepted: 20/09/2026; Published: 30/09/2026

Abstrak– Perkembangan layanan perbankan digital meningkatkan jumlah interaksi pengguna melalui aplikasi *mobile banking*, termasuk BRImo. Ulasan pengguna pada Google Play Store mengandung informasi mengenai pengalaman dan permasalahan layanan yang dapat dianalisis melalui sentimen. Penelitian ini bertujuan menganalisis sentimen ulasan pengguna BRImo serta membandingkan hasil klasifikasi menggunakan metode *Lexicon-Based* dan model BERT. Data penelitian dikumpulkan dari Google Play Store yang menghasilkan 7.000 ulasan. Setelah melalui tahap *cleaning*, *case folding*, *tokenizing*, *stopword removal*, dan *filtering*, diperoleh 4.815 ulasan yang digunakan dalam analisis. Metode *Lexicon-Based* melakukan klasifikasi berdasarkan polaritas kata, sedangkan BERT digunakan sebagai model klasifikasi sentimen berbasis representasi kontekstual. Evaluasi dilakukan menggunakan label referensi berdasarkan *rating* bintang, dengan *rating* 1-2 sebagai negatif, 3 netral, dan 4-5 positif. Hasil menunjukkan *Lexicon-Based* menghasilkan 1.697 sentimen positif, 2.098 netral, dan 1.020 negatif dengan *accuracy* 49,12%. BERT menghasilkan 1.240 positif, 1.639 netral, dan 1.936 negatif dengan *accuracy* 51,11%. BERT juga memperoleh *recall* dan *F1-score* lebih tinggi pada kelas negatif. Hasil penelitian menunjukkan kedua pendekatan dapat digunakan untuk analisis sentimen ulasan BRImo, dengan BERT menghasilkan *accuracy* sedikit lebih tinggi.

Kata Kunci: *Mobile Banking*; analisis sentimen; *Lexicon-Based*; BERT; ulasan pengguna.

Abstract- The development of digital banking services has increased the number of user interactions through mobile banking applications, including BRImo. User reviews on the Google Play Store contain information about user experiences and service-related issues that can be analyzed through sentiment analysis. This study aims to analyze the sentiment of BRImo user reviews and compare the classification results obtained using the Lexicon-Based method and the BERT model. The research data were collected from the Google Play Store, resulting in 7,000 reviews. After the cleaning, case folding, tokenization, stopword removal, and filtering stages, 4,815 reviews were obtained and used in the analysis. The Lexicon-Based method performs classification based on word polarity, while BERT is used as a sentiment classification model based on contextual representations. Evaluation was conducted using reference labels derived from star ratings, with ratings of 1–2 classified as negative, 3 as neutral, and 4–5 as positive. The results show that the Lexicon-Based method produced 1,697 positive, 2,098 neutral, and 1,020 negative sentiments, with an accuracy of 49.12%. BERT produced 1,240 positive, 1,639 neutral, and 1,936 negative sentiments, with an accuracy of 51.11%. BERT also achieved higher recall and F1-score values for the negative class. The results indicate that both approaches can be used for sentiment analysis of BRImo reviews, with BERT achieving slightly higher accuracy.

Keywords: *Mobile Banking*; *Sentiment Analysis*; *Lexicon-Based*; BERT; *User Reviews*.

1. PENDAHULUAN

Transformasi digital telah mendorong perubahan dalam layanan perbankan di Indonesia seiring meningkatnya kebutuhan masyarakat terhadap layanan keuangan yang cepat, praktis, fleksibel, dan dapat diakses kapan saja. Perkembangan teknologi mendorong perbankan menghadirkan berbagai layanan berbasis digital, salah satunya *mobile banking*. Layanan ini memungkinkan nasabah melakukan berbagai aktivitas perbankan melalui *smartphone*, seperti transfer dana, pembayaran tagihan, pembelian produk digital, serta pengecekan informasi rekening tanpa harus datang ke kantor bank atau mesin ATM. Perkembangan tersebut menunjukkan bahwa teknologi digital telah menjadi bagian penting dalam penyediaan layanan perbankan yang sesuai dengan kebutuhan masyarakat modern [1][2]. Pengembangan layanan perbankan digital juga didukung oleh Peraturan Otoritas Jasa Keuangan Nomor 12/POJK.03/2018 tentang Penyelenggaraan Layanan Perbankan Digital oleh Bank Umum yang menekankan penyediaan layanan yang efektif, efisien, mudah, dan aman bagi masyarakat [3][4].

Salah satu layanan *mobile banking* yang digunakan masyarakat Indonesia adalah BRImo, yaitu aplikasi perbankan digital milik Bank Rakyat Indonesia (BRI) yang menyediakan berbagai fasilitas transaksi keuangan. Peningkatan penggunaan aplikasi tersebut juga diikuti dengan bertambahnya ulasan pengguna pada Google Play Store. Ulasan tersebut memuat pengalaman, kritik, keluhan, dan penilaian pengguna terhadap layanan aplikasi sehingga dapat menjadi sumber informasi untuk mengetahui persepsi pengguna. Berbagai keluhan yang disampaikan pengguna, seperti kesalahan aplikasi, kegagalan transaksi, proses login yang lambat, gangguan sistem, hingga aplikasi berhenti secara tiba-tiba, menunjukkan perlunya evaluasi terhadap pengalaman pengguna dalam menggunakan layanan tersebut.

Ulasan pengguna pada Google Play Store merupakan bentuk electronic word of mouth (e-WOM) yang

dapat dimanfaatkan sebagai sumber data untuk mengevaluasi persepsi pengguna. Namun, jumlah ulasan yang besar dan karakteristik data yang tidak terstruktur menyebabkan analisis secara manual membutuhkan waktu serta berpotensi menimbulkan subjektivitas. Salah satu pendekatan yang dapat digunakan untuk mengatasi permasalahan tersebut adalah analisis sentimen [5][6]. Sebagai bagian dari *Natural Language Processing* (NLP), analisis sentimen digunakan untuk mengidentifikasi dan mengklasifikasikan opini dalam data teks ke dalam kategori sentimen positif, negatif, dan netral. Pendekatan ini memungkinkan data ulasan yang tidak terstruktur diolah menjadi informasi yang lebih sistematis mengenai persepsi pengguna terhadap suatu layanan [7][8].

Penelitian terdahulu menunjukkan bahwa analisis sentimen pada aplikasi BRI_{Mo} masih banyak menggunakan metode *machine learning*. Ramadan et al. membandingkan algoritma *Naïve Bayes* dan *Support Vector Machine* (SVM) pada ulasan pengguna BRI_{Mo} dan memperoleh akurasi pengujian SVM sebesar 83,11% [9]. Sementara itu, Nugroho et al. menerapkan model BERT pada ulasan aplikasi *mobile* berbahasa Indonesia dan menunjukkan bahwa BERT mampu memperoleh akurasi hingga 84% pada pendekatan pelabelan berbasis *lexicon* [10]. Berdasarkan penelitian tersebut, masih terdapat ruang untuk melakukan perbandingan secara langsung antara pendekatan *Lexicon-Based* dan BERT pada ulasan pengguna aplikasi BRI_{Mo}, sehingga penelitian ini dilakukan untuk mengetahui perbedaan kinerja kedua pendekatan tersebut dalam klasifikasi sentimen.

Berdasarkan penelitian tersebut, analisis sentimen pada ulasan BRI_{Mo} telah dilakukan menggunakan berbagai pendekatan. Namun, penelitian yang secara khusus melakukan perbandingan langsung antara metode *Lexicon-Based* dan BERT pada ulasan pengguna BRI_{Mo} masih memiliki ruang untuk dikaji. *Lexicon-Based* memiliki keunggulan dalam penerapan karena menggunakan kamus polaritas tanpa memerlukan proses pelatihan, tetapi pendekatan tersebut memiliki keterbatasan dalam memahami konteks kalimat. Sebaliknya, BERT menggunakan representasi kontekstual dua arah sehingga mampu mempertimbangkan hubungan antarkata dalam suatu kalimat [11][12]. Perbedaan karakteristik tersebut menjadi dasar penting untuk membandingkan kemampuan kedua metode dalam mengklasifikasikan sentimen pengguna.

Penelitian ini menggunakan data ulasan pengguna aplikasi BRI_{Mo} yang diperoleh melalui web scraping dari Google Play Store. Data yang dikumpulkan selanjutnya melalui tahap preprocessing sebelum dilakukan klasifikasi menggunakan metode *Lexicon-Based* dan BERT. Kinerja kedua metode dibandingkan menggunakan confusion matrix berdasarkan metrik *accuracy*, *precision*, *recall*, dan *F1-score* [13][14]. Perbandingan tersebut digunakan untuk mengetahui kemampuan masing-masing metode dalam mengklasifikasikan sentimen ulasan pengguna secara objektif [15]. Berdasarkan permasalahan dan research gap tersebut, penelitian ini bertujuan menganalisis sentimen pengguna terhadap aplikasi BRI_{Mo} dengan membandingkan metode *Lexicon-Based* dan BERT. Hasil penelitian diharapkan dapat memberikan gambaran mengenai persepsi pengguna terhadap layanan BRI_{Mo} serta menjadi bahan evaluasi dalam pengembangan kualitas layanan *mobile banking*.

2. METODOLOGI PENELITIAN

Penelitian ini menggunakan pendekatan kuantitatif dengan desain deskriptif-komparatif untuk menganalisis sentimen pengguna terhadap aplikasi *mobile banking* BRI_{Mo} serta membandingkan hasil klasifikasi menggunakan metode *Lexicon-Based* dan BERT (*Bidirectional Encoder Representations from Transformers*). Penelitian dilaksanakan secara *online research* selama periode Mei hingga Agustus 2026 dengan memanfaatkan data sekunder berupa ulasan pengguna yang tersedia pada *Google Play Store*. Data dikumpulkan secara otomatis menggunakan teknik *web scraping* dengan bahasa pemrograman Python dan *library Google Play Scraper*. Atribut yang diperoleh meliputi *review ID*, nama pengguna, isi ulasan, *rating*, jumlah *thumbs up*, versi aplikasi, dan waktu publikasi ulasan [16]. Sebanyak 7.000 ulasan diperoleh sebagai dataset awal. Data selanjutnya diseleksi dan diproses untuk memastikan kelayakannya sebagai data analisis. Ulasan yang digunakan merupakan data berbahasa Indonesia, relevan dengan pengalaman penggunaan aplikasi, memiliki isi yang dapat dianalisis, serta bukan data *spam*. Proses pengolahan menghasilkan 4.815 ulasan bersih yang digunakan pada tahap analisis sentimen. Pengurangan data terutama terjadi karena adanya data kosong, duplikasi, serta dokumen yang tidak memiliki token valid setelah proses *preprocessing*.

2.1 Tahapan Penelitian

Tahapan penelitian terdiri atas pengumpulan data, *preprocessing*, pelabelan sentimen, klasifikasi menggunakan BERT, dan evaluasi model. Alur tersebut dirancang untuk menghasilkan data teks yang terstruktur sebelum dilakukan klasifikasi serta memungkinkan perbandingan kinerja kedua pendekatan yang digunakan [5].

2.1.1 Data Penelitian

Data penelitian diperoleh dari ulasan pengguna aplikasi BRI_{Mo} pada *Google Play Store*. Pengumpulan dilakukan menggunakan teknik *web scraping* berbantuan Python. Sebanyak 7.000 ulasan dikumpulkan sebagai data awal. Data tersebut kemudian disimpan dalam dataset yang memuat informasi ulasan dan atribut

pendukung, seperti *review ID*, *username*, *content*, *rating*, *thumbs up count*, versi aplikasi, dan waktu publikasi.

2.1.2 Data Preprocessing

Preprocessing dilakukan untuk mengurangi *noise* dan menyeragamkan representasi teks sebelum proses klasifikasi. Tahapan yang diterapkan meliputi *cleaning*, *case folding*, penghapusan data kosong dan duplikat, *tokenizing*, *stopword removal*, dan *filtering* [17][18]. Pada tahap *cleaning*, karakter yang tidak relevan, seperti URL, emoji, simbol, angka, tanda baca, dan karakter khusus dihapus. Selanjutnya, *case folding* diterapkan dengan mengubah seluruh teks menjadi huruf kecil. Data kosong dan duplikat kemudian dihapus untuk menghindari pengulangan informasi. Tahap *tokenizing* memecah teks menjadi unit kata, sedangkan *stopword removal* menghilangkan kata-kata yang kurang memberikan kontribusi terhadap analisis sentimen. Tahap terakhir adalah *filtering* berdasarkan panjang karakter token serta penghapusan token kosong. Rangkaian proses tersebut menghasilkan 4.815 ulasan yang digunakan untuk analisis selanjutnya.

2.1.3 Label Sentimen Lexicon-Based

Pelabelan sentimen dilakukan menggunakan metode *Lexicon-Based* dengan InSet Lexicon sebagai kamus polaritas bahasa Indonesia. Setiap token hasil *preprocessing* dicocokkan dengan kamus untuk memperoleh nilai polaritas. Nilai polaritas seluruh token dalam suatu ulasan kemudian dijumlahkan untuk memperoleh skor sentimen dokumen. Secara matematis, skor sentimen dirumuskan sebagai:

$$S(d) = \sum_{i=1}^n P(w_i) \quad (1)$$

dengan $S(d)$ merupakan skor sentimen dokumen, $P(w_i)$ merupakan nilai polaritas token ke- i , dan n merupakan jumlah token dalam dokumen. Penentuan kelas dilakukan berdasarkan nilai skor. Ulasan dengan $S(d) > 0$ dikategorikan sebagai positif, $S(d) = 0$ sebagai netral, dan $S(d) < 0$ sebagai negatif. Hasil pelabelan terhadap 4.815 ulasan menghasilkan 1.697 data positif, 1.020 data negatif, dan 2.098 data netral.

2.1.4 Klasifikasi BERT

Klasifikasi selanjutnya dilakukan menggunakan BERT untuk mempelajari pola hubungan antara teks ulasan dan label sentimen. Dataset yang telah melalui *preprocessing* dan pelabelan digunakan sebagai masukan model. Data dibagi menjadi 80% data pelatihan sebanyak 3.852 ulasan dan 20% data pengujian sebanyak 963 ulasan. Data pelatihan digunakan untuk proses *fine-tuning*, sedangkan data pengujian digunakan untuk mengevaluasi kemampuan model dalam mengklasifikasikan data yang tidak digunakan selama pelatihan.

Pada tahap tokenisasi, teks diproses menggunakan *BERT Tokenizer* berbasis WordPiece untuk mengubah teks menjadi representasi token yang dapat diproses oleh model. Token khusus [CLS] dan [SEP] digunakan sebagai penanda awal dan akhir masukan. Representasi numerik berupa *input IDs* dan *attention mask* kemudian digunakan sebagai masukan model [19].

Model BERT menggunakan arsitektur *Transformer Encoder* dengan mekanisme *self-attention* untuk mempelajari hubungan kontekstual antarkata. Representasi token [CLS] digunakan sebagai representasi keseluruhan teks dan diteruskan ke lapisan klasifikasi untuk menentukan salah satu dari tiga kelas sentimen, yaitu positif, netral, atau negatif. Proses penyesuaian model dilakukan melalui *fine-tuning* menggunakan *Cross Entropy Loss* dan optimizer AdamW sesuai konfigurasi eksperimen penelitian [20].

2.1.5 Evaluasi Model

Evaluasi dilakukan menggunakan *confusion matrix* dan *classification report* untuk mengukur kinerja klasifikasi. Metrik yang digunakan meliputi *accuracy*, *precision*, *recall*, dan *F1-score*. *Accuracy* digunakan untuk mengukur proporsi prediksi yang benar terhadap keseluruhan data, sedangkan *precision* mengukur ketepatan prediksi suatu kelas. *Recall* menunjukkan kemampuan model dalam mengenali data yang termasuk dalam suatu kelas, sementara *F1-score* digunakan untuk mengukur keseimbangan antara *precision* dan *recall* [21].

Performa metode *Lexicon-Based* dan BERT selanjutnya dianalisis berdasarkan hasil evaluasi masing-masing. Perbandingan dilakukan dengan memperhatikan nilai metrik klasifikasi dan pola kesalahan yang ditunjukkan melalui *confusion matrix* [22][23].

3. HASIL DAN PEMBAHASAN

3.1 Hasil Pengumpulan dan Preprocessing Data

Data penelitian berupa ulasan pengguna aplikasi BRImo yang diperoleh dari Google Play Store. Dataset awal yang berhasil dikumpulkan berjumlah 7.000 ulasan. Data tersebut kemudian melalui tahapan *preprocessing* untuk mengurangi *noise* dan menghasilkan teks yang lebih terstruktur sebelum dilakukan analisis sentimen.

Tahapan yang diterapkan meliputi *cleaning*, *case folding*, penghapusan data kosong dan duplikat, *tokenizing*, *stopword removal*, dan *filtering* [24][25].

Hasil setiap tahapan *preprocessing* ditunjukkan pada Tabel 1.

Tabel 1. Hasil Preprocessing Data Ulasan BRImo

Tahapan	Jumlah Data	Keterangan
Dataset awal	7.000	Data hasil <i>scraping</i> Google Play Store
<i>Cleaning</i>	7.000	Penghapusan URL, emoji, simbol, angka, dan karakter khusus
<i>Case folding</i>	7.000	Konversi teks menjadi huruf kecil
Penghapusan data kosong dan duplikat	4.847	Penghapusan data kosong dan ulasan duplikat
<i>Tokenizing</i>	4.847	Pemecahan kalimat menjadi token
<i>Stopword removal</i>	4.847	Penghapusan kata yang tidak relevan
<i>Filtering</i>	4.815	Penyaringan token berdasarkan kriteria yang ditentukan
Dataset akhir	4.815	Data yang digunakan dalam analisis

Tahap *cleaning* dan *case folding* tidak mengurangi jumlah dokumen karena kedua tahap tersebut hanya melakukan perubahan terhadap isi teks. Pengurangan data terjadi setelah pemeriksaan data kosong dan duplikat, kemudian kembali berkurang pada tahap *filtering*. Dengan demikian, diperoleh 4.815 ulasan yang digunakan pada tahap analisis selanjutnya.

Sebagai contoh, ulasan seperti “BRImo Bagus!!! ” setelah *cleaning* menjadi “brimo bagus”, sedangkan kalimat “LOGIN GAGAL TERUS” setelah *case folding* menjadi “login gagal terus”. Pada tahap *tokenizing*, kalimat tersebut dipecah menjadi token [login, gagal, terus]. Tahapan tersebut menghasilkan bentuk data yang lebih seragam dan lebih mudah diproses oleh metode klasifikasi.

3.2 Hasil Klasifikasi *Lexicon-Based*

Setelah *preprocessing*, dilakukan klasifikasi menggunakan metode *Lexicon-Based*. Metode ini menentukan kecenderungan sentimen berdasarkan polaritas kata yang terdapat dalam teks. Setiap token yang memiliki polaritas positif atau negatif memberikan kontribusi terhadap skor sentimen dokumen. Berdasarkan skor tersebut, ulasan dikategorikan menjadi positif, netral, atau negatif sesuai dengan aturan klasifikasi yang diterapkan pada penelitian.

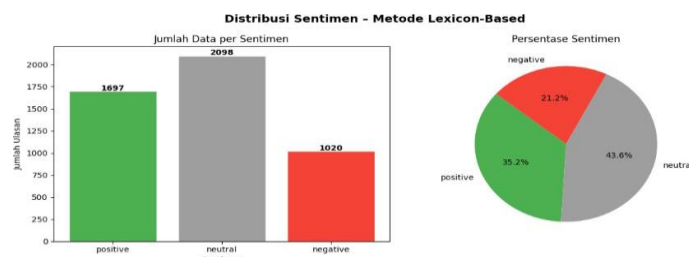
Hasil klasifikasi terhadap 4.815 ulasan ditunjukkan pada Tabel 2.

Tabel 2. Distribusi Sentimen *Lexicon-Based*

Sentimen	Jumlah	Persentas
Positif	1.697	35,25%
Netral	2.098	43,57%
Negatif	1.020	21,18%
Total	4.815	100%

Hasil tersebut menunjukkan bahwa kategori netral memiliki jumlah terbesar, yaitu 2.098 ulasan atau 43,57%. Selanjutnya, sentimen positif berjumlah 1.697 ulasan atau 35,25%, sedangkan sentimen negatif berjumlah 1.020 ulasan atau 21,18%.

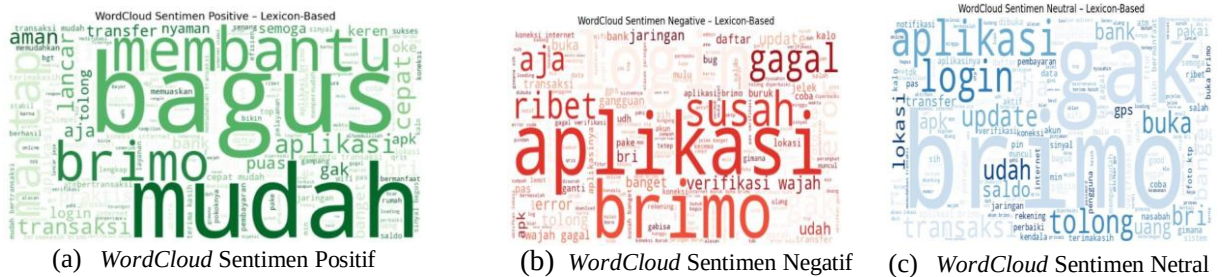
Dominasi kategori netral menunjukkan bahwa pendekatan berbasis kamus dapat menghasilkan kategori netral ketika kata-kata dengan polaritas positif dan negatif muncul secara bersamaan sehingga skor akhirnya relatif seimbang. Kondisi tersebut merupakan salah satu keterbatasan pendekatan *Lexicon-Based* karena klasifikasi terutama bergantung pada polaritas kata dan belum mempertimbangkan hubungan antarkata secara mendalam.



Gambar 1. Distribusi Sentimen *Lexicon-Based*

Visualisasi distribusi sentimen digunakan untuk memperlihatkan proporsi masing-masing kategori secara lebih jelas.

Selain distribusi sentimen, dilakukan visualisasi *WordCloud* untuk melihat kata-kata yang dominan pada setiap kategori sentimen. Visualisasi ini membantu memberikan gambaran mengenai istilah yang sering muncul dalam ulasan pengguna.

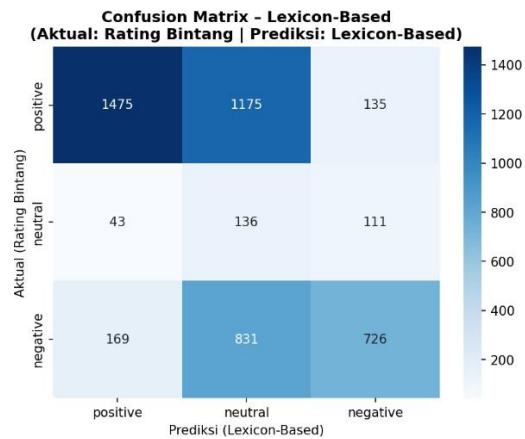


Gambar 2. Wordcloud sentimen (a) Positif, (b) Negatif dan (c) Netral

Hasil *WordCloud* memperlihatkan karakteristik kosakata yang berbeda pada setiap kategori sentimen. Kata-kata yang sering muncul pada masing-masing kategori memberikan gambaran mengenai topik dan pengalaman pengguna yang tercermin dalam ulasan aplikasi BRIMO.

3.3 Evaluasi Metode *Lexicon-Based*

Evaluasi hasil klasifikasi dilakukan menggunakan *Confusion Matrix* dan *Classification Report*. *Confusion Matrix* digunakan untuk melihat kesesuaian antara hasil prediksi dengan label acuan pada masing-masing kelas sentimen.



Gambar 3. Confusion Matrix Lexicon-Based

Evaluasi selanjutnya menggunakan *Precision*, *Recall*, *F1-Score*, dan *Accuracy*. Rumus *Accuracy* yang digunakan adalah:

$$Accuracy = \frac{\text{Jumlah prediksi benar}}{\text{Jumlah seluruh data}} \times 100\%$$

Untuk *Precision*, *Recall*, dan *F1-Score* digunakan: TP

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Hasil evaluasi *Lexicon-Based* disajikan pada Tabel 3.

Tabel 3. Hasil Evaluasi *Lexicon-Based*

Kelas	Precision	Recall	F1-Score
Positif	0,86	0,53	0,65
Netral	0,07	0,46	0,11
Negatif	0,75	0,44	0,55
Accuracy	49,12%		

Nilai *Accuracy* yang diperoleh adalah 49,12%. Pada kelas positif, *Precision* mencapai 0,86 dengan *Recall* sebesar 0,53 dan *F1-Score* sebesar 0,65. Kelas negatif memperoleh *Precision* sebesar 0,75, *Recall* 0,44, dan *F1-Score* 0,55. Sementara itu, kelas netral memiliki nilai *Precision* dan *F1-Score* yang relatif rendah, yaitu masing-masing 0,07 dan 0,11.

Rendahnya performa pada kelas netral menunjukkan adanya kesulitan dalam membedakan ulasan yang benar-benar netral dengan ulasan yang mengandung kecenderungan positif atau negatif. Hal tersebut dapat terjadi karena metode berbasis kamus hanya mempertimbangkan akumulasi polaritas kata dan belum mampu memahami konteks, negasi, maupun kombinasi opini dalam satu kalimat.

3.4 Hasil Klasifikasi Menggunakan BERT

Tahap selanjutnya dilakukan klasifikasi menggunakan model BERT. Model yang digunakan merupakan model bahasa berbasis *Transformer* yang dirancang untuk memproses hubungan antarkata berdasarkan konteks kalimat.

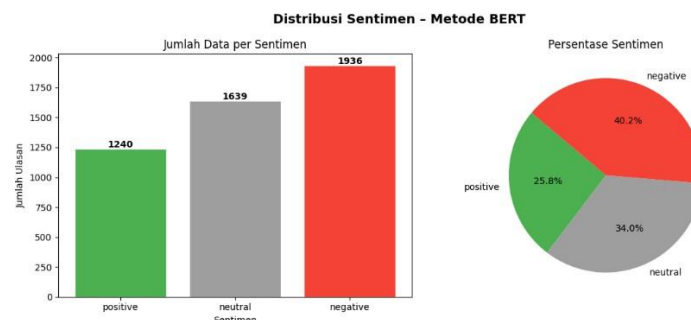
Berbeda dengan *Lexicon-Based*, pendekatan BERT memanfaatkan representasi kontekstual sehingga makna sebuah kata dapat dipengaruhi oleh kata-kata lain yang berada dalam kalimat yang sama. Dalam penelitian ini, model digunakan untuk mengelompokkan ulasan ke dalam tiga kelas, yaitu positif, netral, dan negatif.

Hasil klasifikasi BERT terhadap 4.815 ulasan ditunjukkan pada Tabel 4.

Tabel 4. Distribusi Sentimen BERT

Sentimen	Jumlah	Persentas
Positif	1.240	25,75%
Netral	1.639	34,04%
Negatif	1.936	40,21%
Total	4.815	100%

Berdasarkan hasil tersebut, sentimen negatif menjadi kategori dengan jumlah terbesar, yaitu 1.936 ulasan atau 40,21%. Sentimen netral berjumlah 1.639 ulasan atau 34,04%, sedangkan sentimen positif berjumlah 1.240 ulasan atau 25,75%.



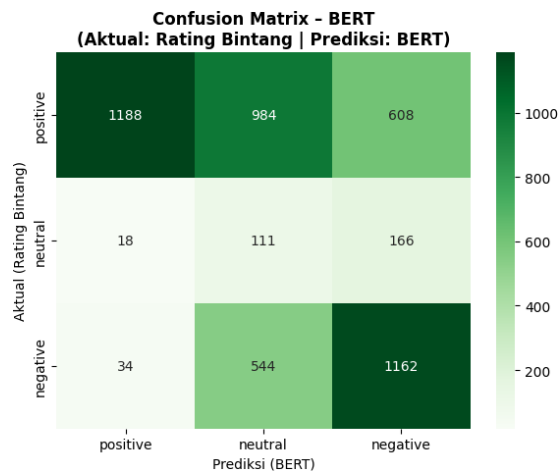
Gambar 4. Distribusi Sentimen BERT

Perbedaan distribusi antara *Lexicon-Based* dan BERT terlihat cukup jelas. Pada *Lexicon-Based*, kategori netral merupakan kategori terbesar, sedangkan pada BERT kategori negatif menjadi kategori dengan jumlah terbesar.

Perbedaan tersebut dapat dikaitkan dengan karakteristik kedua pendekatan. *Lexicon-Based* menentukan sentimen berdasarkan akumulasi polaritas kata, sedangkan BERT mempertimbangkan hubungan antarkata dalam konteks kalimat. Sebagai ilustrasi, kalimat yang mengandung kata positif dan negatif secara bersamaan dapat memperoleh skor yang relatif seimbang pada metode berbasis kamus, sedangkan BERT dapat mempertimbangkan struktur kalimat untuk menentukan kecenderungan makna secara keseluruhan.

3.5 Evaluasi Model BERT

Evaluasi model BERT dilakukan menggunakan *Confusion Matrix* dan *Classification Report*.



Gambar 5. Confusion Matrix BERT

Hasil evaluasi BERT ditunjukkan pada Tabel 5.

Tabel 5. Hasil Evaluasi BERT			
Kelas	Precision	Recall	F1-Score
Positif	0,96	0,43	0,59
Netral	0,07	0,38	0,11
Negatif	0,60	0,67	0,63
Accuracy	51,11%		

BERT memperoleh *Accuracy* sebesar **51,11%**. Nilai *Precision* tertinggi terdapat pada kelas positif, yaitu 0,96. Sementara itu, *Recall* dan *F1-Score* tertinggi terdapat pada kelas negatif, masing-masing sebesar 0,67 dan 0,63.

Sebagai contoh perhitungan *F1-Score* pada kelas negatif:

$$F1 = 2 \times \frac{0,60 \times 0,67}{0,60 + 0,67}$$

$$F1 = \frac{0,804}{1,27} \approx 0,63$$

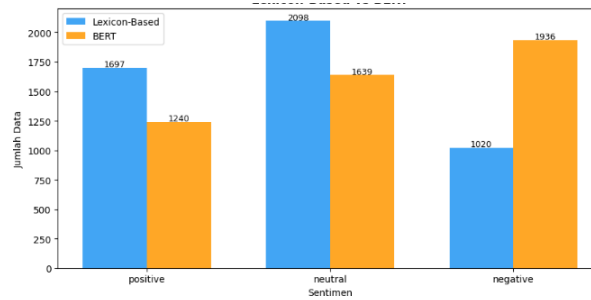
Hasil tersebut menunjukkan bahwa model mampu mempertahankan keseimbangan antara *Precision* dan *Recall* yang lebih baik pada kelas negatif dibandingkan kelas lainnya. Sebaliknya, kelas netral memperoleh *F1-Score* sebesar 0,11 sehingga masih terdapat kesulitan dalam mengenali ulasan netral secara tepat.

3.6 Perbandingan Hasil *Lexicon-Based* dan BERT

Perbandingan dilakukan berdasarkan distribusi sentimen serta metrik evaluasi dari kedua pendekatan.

Tabel 6. Perbandingan Distribusi Sentimen			
Sentimen	<i>Lexicon-Based</i>	BERT	Selisih
Positif	1.697	1.240	-457
Netral	2.098	1.639	-459
Negatif	1.020	1.936	+916
Total	4.815	4.815	

Berdasarkan Tabel 6, terdapat perubahan distribusi yang cukup besar. Jumlah ulasan negatif meningkat dari 1.020 pada *Lexicon-Based* menjadi 1.936 pada BERT. Sebaliknya, jumlah ulasan netral menurun dari 2.098 menjadi 1.639, sedangkan ulasan positif menurun dari 1.697 menjadi 1.240.



Gambar 6. Perbandingan Distribusi Sentimen *Lexicon-Based* dan BERT

Selanjutnya, perbandingan metrik evaluasi ditunjukkan pada Tabel 7.

Tabel 7. Perbandingan Metrik Evaluasi

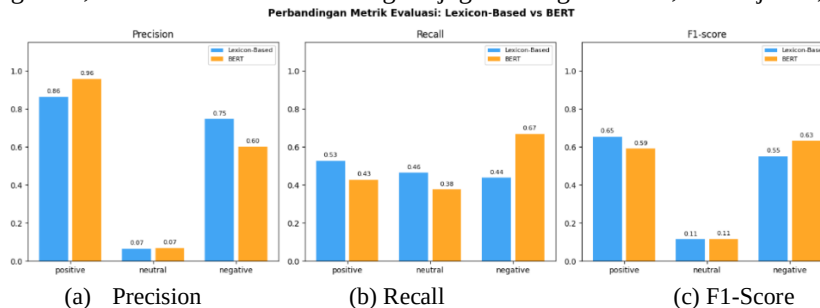
Kelas	Precision LB	Precision BERT	Recall LB	Recall BERT	F1 LB	F1 BERT
Positif	0,86	0,96	0,53	0,43	0,65	0,59
Netral	0,07	0,07	0,46	0,38	0,11	0,11
Negatif	0,75	0,60	0,44	0,67	0,55	0,63

Tabel 8. Perbandingan Accuracy

Metode	Accuracy
<i>Lexicon-Based</i>	49,12%
BERT	51,11%

Hasil evaluasi menunjukkan bahwa *Accuracy* BERT sebesar 51,11%, sedangkan *Lexicon-Based* sebesar 49,12%, sehingga terdapat selisih sebesar 1,99 poin persentase.

Jika dilihat berdasarkan kelas, BERT memiliki *Precision* lebih tinggi pada kelas positif, yaitu 0,96 dibandingkan 0,86 pada *Lexicon-Based*. Namun, *Lexicon-Based* memperoleh *Precision* lebih tinggi pada kelas negatif, yaitu 0,75 dibandingkan 0,60. Pada sisi *Recall*, BERT memperoleh nilai lebih tinggi pada kelas negatif, yaitu 0,67 dibandingkan 0,44. Nilai *F1-Score* kelas negatif juga meningkat dari 0,55 menjadi 0,63.



Gambar 7. Perbandingan Precision, Recall, dan F1-Score

Hasil penelitian memperlihatkan bahwa kedua pendekatan menghasilkan pola klasifikasi yang berbeda terhadap dataset ulasan BRImo. *Lexicon-Based* menghasilkan jumlah sentimen netral paling tinggi, sedangkan BERT menghasilkan jumlah sentimen negatif paling tinggi. Perbedaan tersebut berkaitan dengan mekanisme klasifikasi yang digunakan.

Pada pendekatan *Lexicon-Based*, keputusan sentimen ditentukan berdasarkan keberadaan dan akumulasi kata yang memiliki polaritas tertentu. Pendekatan ini relatif sederhana karena tidak memerlukan proses pembelajaran model, tetapi memiliki keterbatasan ketika menghadapi kalimat yang mengandung negasi, kombinasi opini, bahasa informal, atau konteks yang tidak dapat direpresentasikan hanya melalui polaritas kata.

Sebaliknya, BERT menggunakan representasi kontekstual melalui mekanisme *self-attention*. Setiap token diproses dengan mempertimbangkan hubungan dengan token lainnya dalam suatu kalimat. Mekanisme tersebut memungkinkan model mempertimbangkan konteks secara lebih menyeluruh dalam menghasilkan prediksi.

Perbedaan tersebut terlihat pada hasil evaluasi. BERT memperoleh *Accuracy* sebesar 51,11%, sedangkan *Lexicon-Based* memperoleh 49,12%. Namun, selisih tersebut relatif kecil, yaitu 1,99 poin persentase. Oleh karena itu, hasil penelitian tidak menunjukkan perbedaan performa yang sangat besar secara keseluruhan.

Pada kelas negatif, BERT memperoleh *Recall* 0,67 dan *F1-Score* 0,63, sedangkan *Lexicon-Based* memperoleh masing-masing 0,44 dan 0,55. Hal tersebut menunjukkan bahwa BERT mampu menangkap lebih banyak ulasan yang termasuk dalam kategori negatif berdasarkan label acuan yang digunakan dalam evaluasi. Namun, performa pada kelas netral masih rendah pada kedua pendekatan. Nilai *F1-Score* sebesar 0,11 pada kedua metode menunjukkan bahwa klasifikasi ulasan netral masih menjadi tantangan. Kondisi ini dapat dipengaruhi oleh karakteristik ulasan pengguna yang cenderung singkat, menggunakan bahasa tidak baku, memiliki makna ambigu, atau mengandung lebih dari satu opini dalam satu kalimat.

Dengan demikian, hasil penelitian menunjukkan bahwa BERT menghasilkan Accuracy yang lebih tinggi sebesar 1,99 poin persentase dibandingkan *Lexicon-Based*, sementara karakteristik performa pada setiap kelas berbeda. *Lexicon-Based* memiliki keunggulan pada beberapa metrik kelas tertentu, sedangkan BERT menunjukkan nilai *Recall* dan *F1-Score* yang lebih tinggi pada kelas negatif. Temuan ini menunjukkan bahwa evaluasi analisis sentimen sebaiknya tidak hanya didasarkan pada *Accuracy*, tetapi juga mempertimbangkan *Precision*, *Recall*, dan *F1-Score* pada masing-masing kelas.

4. KESIMPULAN

Berdasarkan hasil penelitian analisis sentimen terhadap ulasan pengguna aplikasi BRImo, tahapan preprocessing yang meliputi cleaning, case folding, tokenizing, stopword removal, dan filtering berhasil menghasilkan data yang lebih bersih dan terstruktur. Dari 7.000 ulasan yang dikumpulkan, diperoleh 4.815 data yang dapat digunakan pada tahap klasifikasi setelah dilakukan penghapusan data kosong, duplikat, dan penyaringan data. Metode *Lexicon-Based* berhasil mengklasifikasikan 4.815 ulasan ke dalam tiga kategori sentimen, yaitu 1.697 positif, 2.098 netral, dan 1.020 negatif, dengan accuracy sebesar 49,12%. Sementara itu, model BERT menghasilkan 1.240 ulasan positif, 1.639 netral, dan 1.936 negatif dengan accuracy sebesar 51,11%. Dengan demikian, terdapat selisih accuracy sebesar 1,99 poin persentase. Hasil evaluasi menunjukkan bahwa BERT memperoleh recall dan F1-score yang lebih tinggi pada kelas negatif, masing-masing sebesar 0,67 dan 0,63, berdasarkan label referensi yang diperoleh dari rating bintang pengguna. Secara keseluruhan, kedua pendekatan dapat digunakan untuk mengidentifikasi kecenderungan sentimen pada ulasan pengguna BRImo, dengan BERT menghasilkan accuracy yang sedikit lebih tinggi dibandingkan *Lexicon-Based*. Hasil analisis juga menunjukkan adanya kata-kata dominan pada ulasan negatif, seperti error, login, OTP, dan transaksi, yang dapat menjadi informasi pendukung dalam evaluasi kualitas layanan dan pengembangan aplikasi. Namun, rendahnya nilai F1-score pada kelas netral menunjukkan bahwa klasifikasi pada kategori tersebut masih menjadi tantangan dan dapat menjadi perhatian untuk penelitian selanjutnya.

REFERENSI

- [1] S. Syuryadi and N. Nurani, "Sistem Informasi Monitoring Komoditas Harga Pangan Berbasis Web Dab Android," *J. It*, vol. 13, no. 3, pp. 77–81, 2023, doi: 10.37639/jti.v13i3.339.
- [2] Z. Sitorus, I. Syahputra, C. Indra Angkat, and D. Sartika, "Implementation of K-Means Clustering for Inventory Projection," *Int. J. Sci. Technol. Manag.*, vol. 5, no. 3, pp. 673–678, 2024, doi: 10.46729/ijstm.v5i3.856.
- [3] A. Rizky *et al.*, "Deep Q-Network Analysis in Optimizing Data Processing for Decision Making on Fuel Expenditure Finance," vol. 07, no. 02, pp. 80–89, 2025.
- [4] Arifin, Z. Sitorus, M. Iqbal, A. P. Utama Siahaan, and S. Wahyuni, "ANALYSIS OF THE SALES POTENTIAL OF BUMDES PRODUCTS USING THE K-MEANS CLUSTERING," vol. 4, no. 1, 2025.
- [5] M. R. Tanjung, M. Iqbal, and Z. Sitorus, "Analisis Sentimen Google Review terhadap Mutu Kualitas Pendidikan pada Perguruan Tinggi STIE Al-Washliyah Sibolga dengan Metode Lexicon dan Algoritma Naive Bayes-Miftah Rusydi Tanjung *et al* Analisis Sentimen Google Review terhadap Mutu Kualitas Pendidikan," vol. 07, no. 02, pp. 2721–1800, 2025, [Online]. Available: <https://journal.cattleyadf.org/index.php/jatilima/index>
- [6] W. Widia, B. Ginting, H. R. Sitepu, L. Nainggolan, A. S. Purba, and A. Br, "Penerapan Algoritma Decision Tree Data Mining untuk Prediksi Pola Pemberian Kredit pada Koperasi Simpan Pinjam," vol. 4, no. 3, pp. 268–276, 2025, doi: 10.47065/jogtc.v4i3.8336.
- [7] Z. Sitorus, M. Iqbal, D. Nasution, and R. Farta Wijaya, "Penerapan Deep Learning dan Analisis Sentimen terhadap Gap Kompetensi Lulusan Lembaga Pendidikan dan Pelatihan Vokasi terhadap Dunia Kerja dengan Metode Long Short-Term Memory (LSTM)," *Bull. Inf. Technol.*, vol. 6, no. 2, pp. 161–172, 2025, doi: 10.47065/bit.v5i2.2029.
- [8] M. Prasetya, M. Wulandari, and S. A. Nikmah, "Implementasi NLP (Natural Language Processing) Dasar pada Analisis Sentiment Review Spotify," *Stain. (Seminar Nas. Teknol. Sains)*, vol. 3, no. 1, pp. 145–153, 2024.
- [9] H. F. Karim and A. P. Wibowo, "Kinerja Metode Fine-Tuning IndoBERT untuk Klasifikasi Emosi Multi-Kelas pada Teks Informal Bahasa Indonesia," vol. 6, no. 1, pp. 63–74, 2025, doi: 10.47065/bulletincsr.v6i1.850.
- [10] A. Darwin, D. Lestari, and I. Seprina, "Perbandingan Kinerja Naive Bayes, Support Vector Machine dan Random Forest Untuk Analisis Sentimen Aplikasi Brimo," vol. 7, no. 3, pp. 1662–1673, 2025, doi: 10.47065/bits.v7i3.8697.
- [11] N. Afifa, R. E. Saputra, and R. A. Nugrahaeni, "Implementasi NLP Pada Chatbot Layanan Akademik Dengan Algoritma Bert," *e-Proceedings Eng.*, vol. 10, no. 1, pp. 383–387, 2023.
- [12] N. Alfriyanto, B. C. Purnama, and F. K. Hasanah, "Analisis Emosi Terhadap Komentar Video Youtube 'Penyebab Kegagalan Adopsi Sistem Pendidikan Finlandia di Indonesia' Menggunakan Metode Random Forest," pp. 812–827, 2024.
- [13] A. Info, "Analisis Sentimen Pada Ulasan Produk Dengan Metode Natural Language Processing (NLP) (Studi Kasus Zalika Store 88 Shopee)," vol. 17, no. 1, pp. 120–128, 2024.
- [14] R. R. Cahya, A. Ichwani, U. Esa, and U. Jakarta, "KOMPARASI KINERJA MODEL SUPPORT VECTOR MACHINE DAN INDONESIAN BIDIRECTIONAL ENCODER REPRESENTATIONS FROM TRANSFORMERS DALAM KLASIFIKASI SENTIMEN," vol. 4307, no. April, pp. 1965–1972, 2026.
- [15] A. Ernawati, Z. Sitorus, M. Iqbal, and D. Nasution, "Penerapan Data Mining Untuk Klasifikasi Penduduk Miskin Di Kabupaten Labuhanbatu Menggunakan Random Forest Dan K-Nearest Neighbors," *Bull. Inf. Technol.*, vol. 6, no. 2, pp. 71–78, 2025, doi: 10.47065/bit.v5i2.1783.

- [16] N. Nurwanda, N. Suarna, and W. Prihartono, “Penerapan Nlp (Natural Language Processing) Dalam Analisis Sentimen Pengguna Telegram Di Playstore,” *JATI (Jurnal Mhs. Tek. Inform.*, vol. 8, no. 2, pp. 1841–1846, 2024, doi: 10.36040/jati.v8i2.8469.
- [17] C. A. Fatikasari, A. M. Rahmania, L. Laili, R. A. Mega Saputri, and M. Arifin, “Analisis Segmentasi dan Prediksi Pola Pembelian IC Label Gamis Menggunakan Hybrid K-Means Random Forest,” *J. Komput. Teknol. Inf. Sist. Komput.*, vol. 5, no. 1, pp. 355–364, 2026, doi: 10.62712/juktisi.v5i1.1007.
- [18] Wahyu Sejati, Ankur Singh Bist, and Amirsyah Tambunan, “Pengembangan Analisis Sentimen dalam Rekayasa Software Engineering menggunakan tinjauan literatur sistematis,” *J. MENTARI Manajemen, Pendidik. dan Teknol. Inf.*, vol. 2, no. 1, pp. 95–103, 2023, doi: 10.33050/mentari.v2i1.377.
- [19] K. S. Nugroho, A. Y. Sukmadewa, F. A. Bachtar, and N. Yudistira, “BERT Fine-Tuning for Sentiment Analysis on Indonesian Mobile Apps Reviews,” pp. 1–10, 2020.
- [20] K. Latifah, A. Handayanto, and N. L. D. M. S., “Comparative Evaluation of Automatic Labeling and Modeling Strategies for Indonesian Sentiment Analysis : Methodology and Performance Evaluation,” vol. 8, no. 3, pp. 1–16, 2026.
- [21] Y. Fauziah, B. Yuwono, and A. S. Aribowo, “Lexicon Based Sentiment Analysis in Indonesia Languages : A Systematic Literature Review,” vol. 1, no. 1, pp. 364–367, 2021, doi: 10.31098/cset.v1i1.397.
- [22] A. M. V. D. V. Id and E. Bleich, “The advantages of *Lexicon-Based* sentiment analysis in an age of machine learning,” pp. 1–19, 2025, doi: 10.1371/journal.pone.0313092.
- [23] L. Barbaglia, S. Consoli, S. Manzan, L. Tiozzo, and E. Tosetti, “Sentiment analysis of economic text: A lexicon-based approach,” no. October 2024, pp. 125–143, 2025, doi: 10.1111/ecin.13264.
- [24] Z. Sitorus, M. Saputra, S. N. Sofyan, and Susilawati, “Sentiment Analysis of Indonesian Community Towards Electric Motorcycles on Twitter Using Orange Data Mining,” *INFOTECH J.*, vol. 10, no. 1, pp. 108–113, 2024, doi: 10.31949/infotech.v10i1.9374.
- [25] D. F. Sjoraida, B. W. K. Guna, and D. Yudhakusuma, “Analisis Sentimen Film Dirty Vote Menggunakan BERT (Bidirectional Encoder Representations from Transformers),” vol. 8, no. 2, 2024.