

EVALUASI KOMPARATIF MODEL SVM, CNN, BIGRU, HIBRIDA CNN-BIGRU, DAN BIGRU-ATTENTION PADA ULASAN TRAVELOKA HOTEL BINTANG EMPAT DI KOTA PADANG

Feriantano Sundang Pranata

Universitas Negeri Padang

Jalan Prof. Dr. Hamka, Air Tawar, Padang, Sumatera Barat, Indonesia

feriantano@unp.ac.id

ABSTRAK

Ulasan hotel pada platform OTA memuat opini pelanggan yang dapat dimanfaatkan untuk memahami persepsi terhadap layanan hotel. Penelitian ini mengevaluasi enam model klasifikasi sentimen, yaitu *TF-IDF + SVM*, *CNN*, *BiGRU*, *CNN-BiGRU*, *BiGRU-CNN*, dan *BiGRU-Attention*, pada 6.975 ulasan Traveloka dari delapan hotel bintang empat di Kota Padang. Keenam model dipilih untuk merepresentasikan spektrum pendekatan klasifikasi teks, yaitu *TF-IDF + SVM* sebagai baseline pembelajaran mesin berbasis fitur, *CNN* untuk menangkap pola lokal atau frasa penanda sentimen, *BiGRU* untuk memodelkan konteks sekuensial dua arah, *CNN-BiGRU* dan *BiGRU-CNN* untuk mengevaluasi pengaruh urutan pemrosesan fitur lokal dan sekuensial, serta *BiGRU-Attention* untuk menilai kontribusi pembobotan selektif terhadap bagian teks yang informatif. *Dataset* diberi label manual ke dalam tiga kelas sentimen, yaitu negatif, netral, dan positif, kemudian diproses melalui tahap pra-pemrosesan dan pembagian data dengan skema 70:15:15. Evaluasi dilakukan menggunakan *accuracy*, *balanced accuracy*, *precision*, *recall*, *macro-F1*, *weighted-F1*, *MCC*, *confusion matrix*, *training time*, dan *inference time*. Hasil evaluasi menunjukkan bahwa *CNN* memperoleh nilai *macro-F1* tertinggi pada eksperimen ini, yaitu 0,7659, dengan *accuracy* 0,8052, *weighted-F1* 0,8100, dan *MCC* 0,6783. *TF-IDF + SVM* menjadi model paling ringan secara komputasi dengan waktu pelatihan 0,3348 detik dan waktu inferensi 0,0398 detik. Analisis *confusion matrix* menunjukkan bahwa kelas netral menjadi kelas yang paling sulit diklasifikasikan karena karakter ulasan yang ambigu dan memuat opini campuran. Temuan ini mengindikasikan bahwa pemilihan model klasifikasi sentimen perlu mempertimbangkan keseimbangan antara performa prediktif, efisiensi komputasi, distribusi kelas, dan kompleksitas arsitektur.

Kata kunci : *BiGRU-Attention*, *CNN-BiGRU*, *Klasifikasi Sentimen*, *TF-IDF-SVM*, *Ulasan Hotel*, *Traveloka*

1. PENDAHULUAN

Online Travel Agent (OTA) menjadi salah satu saluran digital penting dalam layanan perhotelan karena menyediakan informasi hotel, fasilitas pemesanan, serta ruang bagi pelanggan untuk menyampaikan pengalaman menginap melalui ulasan daring. Ulasan pelanggan pada platform OTA merupakan bentuk *user-generated content* yang merepresentasikan persepsi, pengalaman, dan penilaian pengguna terhadap layanan hotel. Dalam konteks perhotelan, ulasan daring telah digunakan sebagai sumber data untuk memahami kepuasan pelanggan, kecenderungan opini, dan persepsi terhadap layanan hotel [1].

Penelitian pada ulasan hotel Indonesia juga menunjukkan bahwa data ulasan dari platform OTA dapat dimanfaatkan untuk pengembangan sistem analisis sentimen berbasis teks [2].

Ulasan hotel memuat opini pelanggan terhadap berbagai dimensi layanan, seperti kamar, kebersihan, lokasi, fasilitas, pelayanan, harga, dan pengalaman menginap secara umum. Data tersebut berbentuk teks bebas sehingga memerlukan pendekatan komputasional untuk mengubah opini pelanggan menjadi informasi yang dapat dianalisis secara sistematis. Kajian sistematis mengenai analisis sentimen ulasan hotel menunjukkan bahwa penelitian

pada domain ini mencakup variasi tahapan pra-pemrosesan, representasi fitur, level analisis sentimen, model klasifikasi, dan *dataset* yang digunakan [1].

Oleh karena itu, analisis sentimen relevan digunakan untuk mengklasifikasikan kecenderungan opini pelanggan berdasarkan isi teks ulasan hotel [3].

Hotel bintang empat di Kota Padang dipilih sebagai objek penelitian berdasarkan pertimbangan metodologis dan ketersediaan data. Pembatasan pada satu kategori hotel bertujuan mengurangi heterogenitas yang dapat muncul akibat perbedaan kelas layanan, fasilitas, segmen konsumen, dan ekspektasi pelanggan antara hotel berbintang dan non-bintang maupun antar-kategori bintang. Dengan menggunakan kategori hotel yang relatif setara, variasi sentimen dalam ulasan dapat dianalisis dalam konteks layanan yang lebih terkendali. Pembatasan geografis pada Kota Padang juga dilakukan untuk menjaga konsistensi konteks lokasi, karakteristik destinasi, dan lingkungan pelayanan hotel. Selain itu, hotel bintang empat di Kota Padang memiliki volume ulasan Traveloka yang memadai untuk membangun *dataset* tiga kelas dan melakukan evaluasi komparatif terhadap enam model klasifikasi. Pemilihan objek tersebut tidak dimaksudkan untuk merepresentasikan seluruh hotel di Indonesia, tetapi sebagai konteks empiris yang

terkontrol untuk menguji perbedaan performa model pada ulasan hotel berbahasa Indonesia.

Analisis sentimen merupakan pendekatan dalam *natural language processing* yang digunakan untuk mengidentifikasi dan mengklasifikasikan opini dalam teks ke dalam kategori sentimen tertentu, seperti positif, netral, dan negatif. Dalam penelitian ini, klasifikasi sentimen dilakukan melalui evaluasi komparatif terhadap model pembelajaran mesin klasik dan model berbasis jaringan saraf. Model *TF-IDF* + *SVM* digunakan sebagai pembanding awal (*baseline*) karena pendekatan tersebut masih banyak digunakan dalam klasifikasi sentimen berbasis fitur teks [4]. Selain itu, model *CNN*, *BiGRU*, *CNN-BiGRU*, *BiGRU-CNN*, dan *BiGRU-Attention* digunakan untuk merepresentasikan pendekatan jaringan saraf yang umum diterapkan dalam klasifikasi teks dan analisis sentimen [5]–[9].

Pemilihan enam model tersebut bertujuan memperoleh gambaran empiris mengenai perbedaan kinerja beberapa pendekatan klasifikasi sentimen pada data ulasan hotel. *TF-IDF* + *SVM* merepresentasikan pendekatan klasik berbasis fitur, *CNN* dan *BiGRU* merepresentasikan model jaringan saraf tunggal, *CNN-BiGRU* dan *BiGRU-CNN* merepresentasikan variasi model *hybrid* yang menggabungkan pemrosesan lokal dan sekuensial, sedangkan *BiGRU-Attention* merepresentasikan model sekuensial dengan mekanisme perhatian. Dengan demikian, perbandingan model tidak hanya mencakup pendekatan klasik dan jaringan saraf, tetapi juga mencakup variasi arsitektur *hybrid* dan *attention-based* yang relevan untuk klasifikasi sentimen teks. Desain tersebut memungkinkan perbandingan dilakukan tidak hanya berdasarkan performa klasifikasi, tetapi juga berdasarkan pengaruh representasi fitur, pemodelan urutan, urutan penggabungan arsitektur, dan kompleksitas komputasi.

Penelitian terdahulu telah membahas analisis sentimen ulasan hotel menggunakan berbagai pendekatan, baik pembelajaran mesin klasik, *deep learning*, maupun analisis berbasis aspek [2], [3], [10].

Namun, evaluasi komparatif terhadap *TF-IDF* + *SVM*, *CNN*, *BiGRU*, *CNN-BiGRU*, *BiGRU-CNN*, dan *BiGRU-Attention* pada ulasan Traveloka hotel bintang empat di Kota Padang masih perlu dikaji secara lebih spesifik. Sebagian penelitian sebelumnya berfokus pada ulasan hotel Indonesia secara umum, sebagian lain menggunakan pendekatan *multilevel* atau *aspect-based sentiment analysis*, sedangkan penelitian ini difokuskan pada perbandingan enam model klasifikasi sentimen dalam konteks data, lokasi, dan kelas hotel yang lebih spesifik.

Berdasarkan permasalahan tersebut, penelitian ini bertujuan mengevaluasi kinerja model *TF-IDF* + *SVM*, *CNN*, *BiGRU*, *CNN-BiGRU*, *BiGRU-CNN*, dan *BiGRU-Attention* untuk klasifikasi sentimen ulasan Traveloka pada hotel bintang empat di Kota Padang. Kontribusi penelitian ini meliputi penyusunan *dataset*

klasifikasi sentimen berbasis teks dari ulasan Traveloka hotel bintang empat di Kota Padang, evaluasi komparatif terhadap enam model klasifikasi sentimen, serta penggunaan metrik performa dan efisiensi komputasi, seperti *accuracy*, *precision*, *recall*, *macro-F1*, *weighted-F1*, *confusion matrix*, *classification report*, *training time*, dan *inference time*. Dengan demikian, penelitian ini memberikan kontribusi empiris dan eksperimental dalam menilai kesesuaian beberapa model klasifikasi sentimen pada ulasan hotel yang tersedia di platform Traveloka.

2. TINJAUAN PUSTAKA

2.1. Analisis Sentimen

Analisis sentimen merupakan bidang dalam *natural language processing* dan *text mining* yang bertujuan mengidentifikasi, mengekstraksi, dan mengklasifikasikan opini dalam teks ke dalam kategori sentimen tertentu, seperti positif, netral, dan negatif. Dalam domain perhotelan, analisis sentimen digunakan untuk memahami persepsi pelanggan berdasarkan pengalaman yang ditulis dalam ulasan daring, sehingga dapat membantu mengidentifikasi kepuasan, keluhan, dan penilaian pelanggan terhadap layanan hotel. Kajian sistematis mengenai analisis sentimen ulasan hotel menunjukkan bahwa penelitian pada domain ini mencakup variasi teknik pra-pemrosesan, representasi fitur, level klasifikasi sentimen, model klasifikasi, dan *dataset* yang digunakan [1].

Dengan demikian, analisis sentimen relevan digunakan sebagai dasar untuk mengklasifikasikan kecenderungan opini pelanggan pada ulasan hotel.

2.2. TF-IDF dan SVM

Term Frequency-Inverse Document Frequency (*TF-IDF*) merupakan teknik representasi teks yang memberikan bobot pada kata berdasarkan frekuensi kemunculannya dalam dokumen dan tingkat kelangkaannya dalam korpus [11].

Representasi ini banyak digunakan dalam klasifikasi teks karena mampu mengubah teks menjadi vektor numerik yang dapat diproses oleh algoritma pembelajaran mesin. *Support Vector Machine* (*SVM*) merupakan algoritma klasifikasi berbasis margin yang bertujuan mencari bidang pemisah optimal antar kelas [12].

Dalam klasifikasi sentimen berbasis teks, kombinasi *TF-IDF* + *SVM* masih relevan digunakan sebagai pembanding awal karena sesuai untuk fitur teks berdimensi tinggi dan telah digunakan pada analisis sentimen ulasan hotel berbahasa Indonesia [4].

Oleh karena itu, penelitian ini menggunakan *TF-IDF* + *SVM* sebagai model pembanding terhadap model berbasis jaringan saraf.

2.3. CNN untuk Klasifikasi Teks

Convolutional Neural Network (*CNN*) dalam klasifikasi teks menggunakan operasi konvolusi untuk menangkap pola lokal, frasa pendek, atau n-gram yang

berkontribusi terhadap kelas tertentu. Dalam konteks ulasan hotel, frasa seperti “kamar bersih”, “pelayanan lambat”, “lokasi strategis”, atau “fasilitas kurang” dapat menjadi penanda sentimen yang penting. Model *CNN* untuk klasifikasi kalimat menunjukkan bahwa operasi konvolusi pada representasi kata dapat digunakan untuk mengekstraksi fitur diskriminatif pada tugas klasifikasi teks [5].

Berdasarkan karakteristik tersebut, *CNN* relevan digunakan dalam penelitian ini karena ulasan hotel sering memuat frasa pendek yang merepresentasikan penilaian pelanggan terhadap layanan.

2.4. BiGRU

Gated Recurrent Unit (GRU) merupakan arsitektur *recurrent neural network* yang dirancang untuk memproses data sekuensial dengan mekanisme gerbang yang lebih sederhana dibandingkan *Long Short-Term Memory (LSTM)* [6].

GRU digunakan untuk mengatur informasi yang dipertahankan atau diperbarui selama pemrosesan urutan, sehingga sesuai untuk data teks yang memiliki hubungan antarkata. *Bidirectional GRU (BiGRU)* memanfaatkan prinsip *bidirectional recurrent neural network*, yaitu membaca urutan dari dua arah agar model dapat menangkap konteks sebelum dan sesudah suatu kata [13].

Dalam klasifikasi sentimen ulasan, *BiGRU* relevan digunakan karena polaritas opini tidak hanya dipengaruhi oleh keberadaan kata tertentu, tetapi juga oleh urutan dan konteks kemunculannya dalam kalimat.

2.5. CNN-BiGRU dan BiGRU-CNN

CNN-BiGRU dan *BiGRU-CNN* merupakan variasi model hybrid yang menggabungkan kemampuan *CNN* dalam mengekstraksi fitur lokal dengan kemampuan *BiGRU* dalam memodelkan konteks sekuensial. Pada *CNN-BiGRU*, *CNN* ditempatkan sebelum *BiGRU* sehingga fitur lokal diekstraksi terlebih dahulu sebelum diproses secara sekuensial. Sebaliknya, pada *BiGRU-CNN*, *BiGRU* ditempatkan sebelum *CNN* sehingga representasi sekuensial dua arah dibentuk terlebih dahulu, kemudian *CNN* digunakan untuk mengekstraksi pola lokal dari representasi tersebut. Penelitian mengenai *CNN + BiGRU* pada analisis sentimen berbasis aspek menunjukkan bahwa *CNN* dapat digunakan untuk menangkap fitur lokal, sedangkan *BiGRU* digunakan untuk mempelajari dependensi urutan dalam teks [7].

Model *hybrid* berbasis *CNN-BiGRU* juga dikembangkan dalam klasifikasi sentimen untuk memperkuat representasi fitur melalui penggabungan informasi lokal dan sekuensial [8].

Dalam penelitian ini, kedua variasi tersebut digunakan untuk mengevaluasi pengaruh urutan pemrosesan lokal-sekuensial dan sekuensial-lokal terhadap klasifikasi sentimen ulasan hotel.

2.6. BiGRU-Attention

Mekanisme *attention* merupakan pendekatan yang memberikan bobot berbeda pada bagian *input* berdasarkan kontribusinya terhadap keluaran model. Secara konseptual, *attention* dikembangkan untuk membantu model memfokuskan representasi pada bagian *input* yang lebih relevan terhadap proses prediksi [14].

Pada model *BiGRU-Attention*, *BiGRU* digunakan untuk menghasilkan representasi sekuensial dua arah, sedangkan *attention* digunakan untuk menonjolkan bagian teks yang lebih informatif terhadap keputusan klasifikasi. Dalam ulasan hotel, tidak semua kata memiliki kontribusi yang sama terhadap sentimen; kata atau frasa yang berkaitan dengan keluhan, kepuasan, pelayanan, kebersihan, atau fasilitas dapat memiliki bobot yang lebih tinggi dalam penentuan polaritas. Penelitian mengenai ulasan hotel daring menunjukkan bahwa *BiGRU* dengan mekanisme *attention* dapat digunakan untuk analisis sentimen ulasan hotel, sedangkan model *recurrent* berbasis *attention* juga dikembangkan untuk menangkap kata-kata emosional dalam ulasan hotel [9], [15].

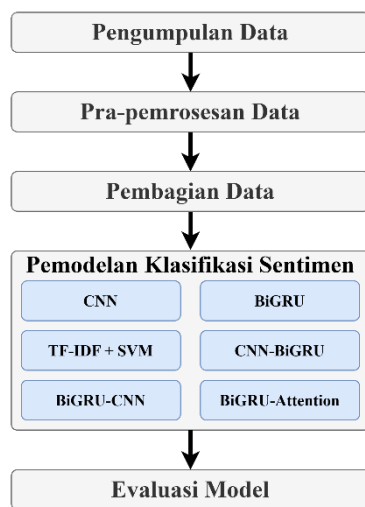
2.7. Posisi Penelitian terhadap Studi Terdahulu

Penelitian terdahulu menunjukkan bahwa analisis sentimen ulasan hotel telah dikaji dengan berbagai pendekatan, konteks data, dan level analisis. Kusumaningrum et al. [3] membandingkan *Logistic Regression*, *Naïve Bayes*, *SVM*, dan *CNN* pada ulasan hotel Traveloka Indonesia dan menunjukkan bahwa pendekatan *CNN* berbasis *Word2Vec* memperoleh kinerja lebih tinggi dibandingkan metode klasik, tetapi penelitian tersebut masih berfokus pada klasifikasi biner dan belum mengevaluasi *BiGRU*, *CNN-BiGRU*, *BiGRU-CNN*, serta *BiGRU-Attention*. Studi lain oleh Kusumaningrum et al. [2] mengembangkan aplikasi analisis sentimen multilevel berbasis *CNN* dan *improved LSTM* pada ulasan hotel Indonesia, tetapi fokusnya diarahkan pada analisis level dokumen dan aspek, bukan evaluasi komparatif model klasifikasi sentimen. Cahyaningtyas et al. [10] membahas *aspect-based sentiment analysis* pada ulasan hotel Indonesia menggunakan model *deep learning*, tetapi ruang lingkupnya lebih menekankan klasifikasi aspek dan polaritas aspek.

Priya dan Deepalakshmi [16] menerapkan teknik *deep learning* pada ulasan hotel dari jejaring sosial, Erdoğan et al. [17] mengembangkan model sentimen berbasis *deep learning* pada ulasan hotel bintang lima, sedangkan Yanuargi et al. [18] menggunakan pendekatan *multistage BiLSTM* pada ulasan hotel *Google Maps* di Yogyakarta. Berdasarkan studi-studi tersebut, penelitian ini memosisikan diri sebagai evaluasi komparatif *TF-IDF + SVM*, *CNN*, *BiGRU*, *CNN-BiGRU*, *BiGRU-CNN*, dan *BiGRU-Attention* pada ulasan Traveloka hotel bintang empat di Kota Padang, dengan evaluasi yang mencakup performa klasifikasi dan efisiensi komputasi.

3. METODE PENELITIAN

Penelitian ini menggunakan desain eksperimen komparatif dengan pendekatan kuantitatif untuk mengevaluasi beberapa model klasifikasi sentimen pada ulasan Traveloka hotel bintang empat di Kota Padang. Model yang dievaluasi terdiri dari *TF-IDF + SVM*, *CNN*, *BiGRU*, *CNN-BiGRU*, *BiGRU-CNN*, dan *BiGRU-Attention*. *Input* utama penelitian berupa teks ulasan pelanggan, sedangkan *output* model berupa kelas sentimen negatif, netral, dan positif. Secara umum, tahapan penelitian meliputi pengumpulan data, pra-pemrosesan data, pembagian data, pemodelan klasifikasi sentimen, pengaturan eksperimen, dan evaluasi model.



Gambar 1. Alur penelitian

3.1. Pengumpulan Data

Data penelitian diperoleh dari ulasan pelanggan hotel yang tersedia pada platform Traveloka dengan ruang lingkup hotel bintang empat yang berlokasi di Kota Padang, Sumatera Barat. Penggunaan ulasan dari platform OTA sebagai sumber data telah digunakan dalam penelitian analisis sentimen ulasan hotel karena ulasan pelanggan memuat pengalaman pengguna terhadap layanan hotel dan dapat dimanfaatkan sebagai sumber data klasifikasi sentimen berbasis teks [2], [3].

Pengumpulan data dilakukan melalui ekstraksi informasi dari struktur *HTML* halaman ulasan hotel pada platform Traveloka. Konten *HTML* yang memuat ulasan pelanggan diekstraksi, kemudian hasil ekstraksi disusun dalam format *tabular* untuk diproses pada tahap berikutnya. Teks yang berasal dari ulasan tamu dipertahankan, sedangkan teks yang merupakan balasan dari pihak hotel tidak digunakan karena unit analisis penelitian dibatasi pada opini pelanggan. Atribut utama yang digunakan meliputi sumber platform, nama hotel, lokasi hotel, kelas bintang hotel, dan teks ulasan pelanggan. Informasi pribadi pengguna tidak digunakan dalam proses analisis.

3.2. Pra-pemrosesan Data

Pra-pemrosesan data dilakukan untuk menyiapkan teks ulasan agar dapat digunakan secara konsisten pada tahap pemodelan. Sebelum proses pemodelan, label sentimen hasil anotasi manual divalidasi dan diseragamkan ke dalam tiga kelas, yaitu negatif, netral, dan positif. Label negatif diberikan pada ulasan yang dominan memuat keluhan atau ketidakpuasan, label positif diberikan pada ulasan yang dominan memuat kepuasan atau penilaian baik, sedangkan label netral diberikan pada ulasan yang bersifat informatif, ambigu, atau memuat opini campuran tanpa kecenderungan sentimen yang dominan. *Rating* ulasan tidak digunakan sebagai label sentimen maupun fitur *input* model karena penelitian berfokus pada klasifikasi sentimen berdasarkan isi teks ulasan. Dalam analisis sentimen ulasan hotel, pra-pemrosesan merupakan tahap penting karena ulasan pelanggan umumnya berbentuk teks bebas dan mengandung variasi penulisan, simbol, *emoji*, pengulangan karakter, singkatan, serta kata tidak baku [2].

Tahapan pra-pemrosesan yang digunakan meliputi pembersihan teks, *case folding*, normalisasi kata tidak baku, pemetaan *emoji*, normalisasi istilah khusus, pengurangan karakter berulang, penghapusan karakter tidak relevan, *tokenisasi*, dan *padding*, sebagaimana ditunjukkan pada Tabel 1.

Tabel 1. Tahapan pra-pemrosesan data ulasan

Tahap	Prosedur	Tujuan	Contoh
Pembersihan teks	Menghapus URL dan elemen yang tidak relevan	Mengurangi gangguan pada teks	www... kamar bagus → kamar bagus
<i>Case folding</i>	Mengubah seluruh huruf menjadi huruf kecil	Menyeragamkan bentuk token	Kamar Bersih → kamar bersih
Normalisasi kata	Mengubah singkatan dan kata tidak baku menjadi bentuk lebih baku	Mengurangi variasi penulisan	tdk nyaman → tidak nyaman
Pemetaan <i>emoji</i>	Mengubah <i>emoji</i> bermuatan sentimen menjadi token tekstual	Mempertahankan sinyal sentimen dari <i>emoji</i>	😊 → emoji_positif
Normalisasi istilah khusus	Menyeragamkan istilah dalam konteks hotel	Menstabilkan istilah domain	check in → checkin
Pengurangan karakter berulang	Mengurangi pengulangan karakter berlebihan menjadi dua karakter	Mengurangi variasi token yang terlalu ekstrem	lamaaa → lamaa
Penghapusan karakter tidak relevan	Menghapus tanda baca dan simbol yang tidak diperlukan	Menstabilkan token teks	bagus!!! → bagus

Tahap	Prosedur	Tujuan	Contoh
Penghapusan spasi berlebih	Menyeragamkan jarak antarkata	Membentuk teks bersih	kamar bersih → kamar bersih
Tokenisasi	Mengubah teks menjadi urutan token	Menyiapkan teks untuk representasi numerik	kamar bersih → [kamar, bersih]
Padding	Menyamakan panjang urutan token	Menyiapkan input berdimensi seragam	[kamar, bersih] → [kamar, bersih, PAD]

3.3. Pembagian Data

Setelah tahap pra-pemrosesan teks, *dataset* dibagi menjadi tiga *subset*, yaitu data latih, data validasi, dan data uji. Data latih digunakan untuk membangun model, data validasi digunakan untuk memantau proses pelatihan dan pemilihan model selama pembelajaran, sedangkan data uji digunakan untuk mengevaluasi performa akhir model pada data yang tidak terlibat dalam proses pelatihan. Pembagian dilakukan dengan skema 70% data latih, 15% data validasi, dan 15% data uji. Proses pembagian menggunakan pendekatan *stratified split* agar proporsi kelas sentimen pada setiap *subset* tetap merepresentasikan distribusi kelas pada *dataset*. Selain itu, pembobotan kelas diterapkan pada model yang mendukung mekanisme tersebut untuk mengurangi potensi bias terhadap kelas mayoritas, karena distribusi kelas yang tidak seimbang dapat menyebabkan model cenderung menghasilkan prediksi yang menguntungkan kelas dominan [19]. Bobot kelas dihitung menggunakan Persamaan (1).

$$w_c = \frac{N}{K \times N_c} \quad (1)$$

dengan w_c adalah bobot untuk kelas ke- c , N adalah jumlah seluruh data latih, K adalah jumlah kelas sentimen, dan N_c adalah jumlah data latih pada kelas ke- c .

3.4. Pemodelan Klasifikasi Sentimen

Pemodelan klasifikasi sentimen dilakukan untuk membandingkan enam model dalam mengklasifikasikan teks ulasan ke dalam kelas sentimen negatif, netral, dan positif. Model yang digunakan terdiri dari *TF-IDF + SVM*, *CNN*, *BiGRU*, *CNN-BiGRU*, *BiGRU-CNN*, dan *BiGRU-Attention*. Secara umum, skema pemodelan dibedakan menjadi dua pendekatan representasi. Pendekatan pertama menggunakan representasi *TF-IDF* sebagai masukan untuk *SVM*, sedangkan pendekatan kedua menggunakan *tokenisasi* dan *embedding* sebagai representasi awal pada model berbasis jaringan saraf. Alur pemodelan klasifikasi sentimen dalam penelitian ini ditunjukkan pada Gambar 2.

Model *TF-IDF + SVM* digunakan sebagai pembanding awal terhadap model berbasis jaringan saraf. Pada skema ini, teks ulasan hasil pra-pemrosesan direpresentasikan ke dalam bentuk vektor berbobot menggunakan *TF-IDF*. Pembobotan *TF-IDF* mempertimbangkan frekuensi kemunculan term dalam dokumen dan tingkat kelangkaannya dalam korpus [11].

Representasi vektor tersebut kemudian digunakan sebagai masukan bagi *SVM* untuk

memisahkan data ke dalam kelas sentimen berdasarkan karakteristik fitur teks. *SVM* digunakan sebagai *classifier* berbasis margin yang bertujuan membentuk bidang pemisah optimal antar kelas [12].

Arsitektur *CNN* digunakan untuk mempelajari pola lokal dalam teks ulasan. Teks yang telah dikonversi menjadi urutan token direpresentasikan melalui lapisan *embedding*, kemudian diproses menggunakan konvolusi satu dimensi. Mekanisme konvolusi memungkinkan model menangkap fitur lokal, seperti frasa pendek atau pola *n-gram* yang berkaitan dengan kelas sentimen tertentu. Penggunaan *CNN* pada klasifikasi kalimat menunjukkan bahwa operasi konvolusi dapat mengekstraksi fitur lokal dari representasi kata secara efektif [5].

Model *BiGRU* digunakan untuk mempelajari informasi sekuensial dalam teks ulasan. Setelah teks direpresentasikan melalui *tokenisasi* dan *embedding*, urutan token diproses menggunakan *Bidirectional GRU*. *GRU* memproses data sekuensial melalui mekanisme gerbang yang mengatur informasi yang dipertahankan dan diperbarui selama pembacaan urutan [6].

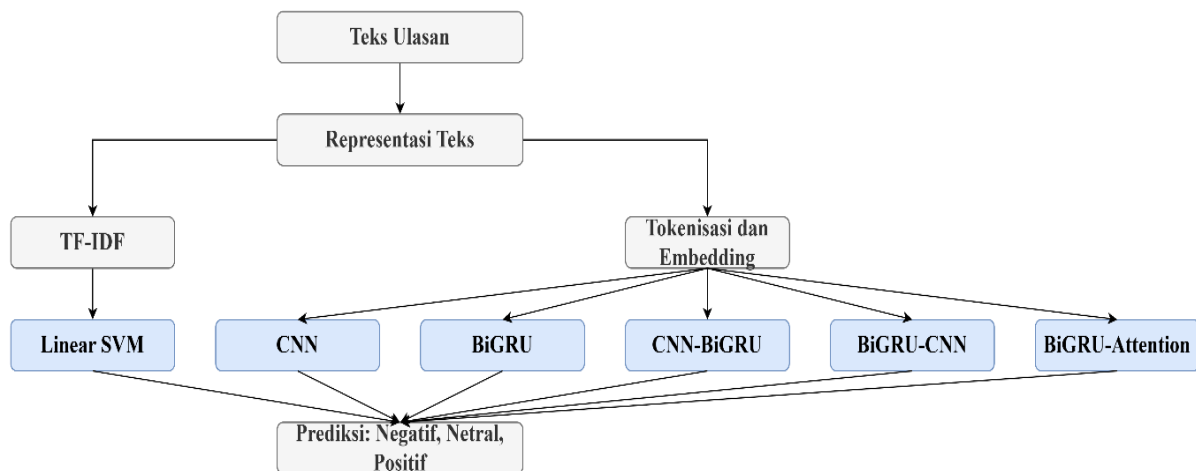
Sementara itu, struktur *bidirectional* memungkinkan model membaca teks dari dua arah, yaitu dari awal ke akhir dan dari akhir ke awal, sehingga konteks sebelum dan sesudah suatu kata dapat digunakan dalam pembentukan representasi [13].

Dua varian arsitektur *hybrid*, yaitu *CNN-BiGRU* dan *BiGRU-CNN*, digunakan untuk mengevaluasi pengaruh urutan penggabungan komponen konvolusional dan sekuensial. *CNN-BiGRU* menempatkan *CNN* sebelum *BiGRU*, sehingga fitur lokal diekstraksi terlebih dahulu sebelum diproses secara sekuensial. Sebaliknya, *BiGRU-CNN* menempatkan *BiGRU* sebelum *CNN*, sehingga representasi sekuensial dua arah dibentuk terlebih dahulu sebelum pola lokal diekstraksi dari keluaran sekuensial. Perbandingan kedua varian ini digunakan untuk menilai apakah urutan pemrosesan lokal-sekuensial dan sekuensial-lokal menghasilkan perbedaan kinerja pada klasifikasi sentimen ulasan hotel. Penggunaan kombinasi *CNN* dan *BiGRU* mengacu pada pendekatan yang memanfaatkan informasi lokal dan informasi urutan dalam teks [7], [8].

Model *BiGRU-Attention* digunakan untuk mengevaluasi kontribusi mekanisme perhatian dalam klasifikasi sentimen. Arsitektur ini memanfaatkan *BiGRU* untuk membentuk representasi sekuensial dua arah, kemudian menerapkan mekanisme *attention* untuk memberikan bobot berbeda pada bagian teks

berdasarkan kontribusinya terhadap keputusan klasifikasi. Mekanisme *attention* membantu model memfokuskan representasi pada bagian *input* yang lebih relevan terhadap keluaran [14]. Dengan

demikian, *BiGRU-Attention* digunakan untuk menilai apakah pembobotan terhadap kata atau frasa yang lebih informatif dapat mendukung klasifikasi sentimen ulasan hotel.



Gambar 2. Alur pemodelan klasifikasi sentimen

3.5. Pengaturan Eksperimen

Eksperimen dilakukan menggunakan *Google Colab* dengan akselerator *GPU NVIDIA T4*. Model *TF-IDF + SVM* diimplementasikan menggunakan pustaka *scikit-learn*, sedangkan model *CNN*, *BiGRU*, *CNN-BiGRU*, *BiGRU-CNN*, dan *BiGRU-Attention* diimplementasikan menggunakan *TensorFlow/Keras*. Seluruh model dilatih dan diuji menggunakan skema pembagian data yang sama agar hasil evaluasi dapat dibandingkan secara konsisten. Nilai acak ditetapkan sebesar 42 untuk menjaga konsistensi pembagian data dan proses pelatihan. Pada model *TF-IDF + SVM*, representasi teks menggunakan *unigram* dan *bigram* dengan pembatasan jumlah fitur, penyaringan term yang terlalu jarang atau terlalu umum, serta pembobotan frekuensi secara *sublinear*. Model *SVM* yang digunakan adalah *SVM linear* dengan pembobotan kelas untuk menangani ketidakseimbangan distribusi sentimen.

Pada model berbasis jaringan saraf, teks ulasan direpresentasikan sebagai urutan *token* numerik dengan batas kosakata sebanyak 20.000 kata dan panjang maksimum 120 *token*. Representasi kata dipelajari melalui lapisan *embedding* berdimensi 128. Model yang menggunakan komponen *GRU* menggunakan 64 unit *GRU*, sedangkan model berbasis *CNN* menggunakan 128 filter konvolusi dengan ukuran *kernel* 5. Proses pelatihan model berbasis jaringan saraf menggunakan *optimizer Adam* dengan laju pembelajaran awal 3×10^{-4} , ukuran batch 32, dan jumlah *epoch* maksimum 30. *Adam* digunakan karena mampu memperbarui bobot secara adaptif berdasarkan estimasi momen pertama dan kedua dari gradien [20].

Fungsi *loss* yang digunakan adalah *sparse categorical cross-entropy* karena label sentimen direpresentasikan dalam bentuk indeks kelas,

sedangkan lapisan keluaran menggunakan fungsi aktivasi *softmax* untuk menghasilkan probabilitas pada setiap kelas sentimen. Untuk mengurangi risiko *overfitting*, pelatihan menerapkan *dropout* sebesar 0,40, *spatial dropout* sebesar 0,30, *dropout* pada komponen *GRU* sebesar 0,30, serta regularisasi *L2* sebesar 1×10^{-4} . *Dropout* digunakan untuk menonaktifkan sebagian unit secara acak selama pelatihan sehingga ketergantungan berlebih antarunit jaringan dapat dikurangi [21].

Selain itu, *early stopping* diterapkan berdasarkan *validation loss* dengan toleransi 3 *epoch*, penurunan laju pembelajaran diterapkan ketika *validation loss* tidak membaik dalam 2 *epoch*, dan model terbaik ditentukan berdasarkan nilai *validation loss* terendah selama proses pelatihan.

3.6. Evaluasi Model

Evaluasi model dilakukan untuk mengukur kinerja setiap model klasifikasi sentimen pada data uji. Metrik evaluasi yang digunakan meliputi *accuracy*, *precision*, *recall*, *F1-score*, *macro-F1*, *weighted-F1*, *balanced accuracy*, *Matthews Correlation Coefficient (MCC)*, *confusion matrix*, *normalized confusion matrix*, *classification report*, *training time*, dan *inference time*. Penggunaan beberapa metrik diperlukan karena *accuracy* saja tidak selalu cukup untuk mengevaluasi model klasifikasi, terutama ketika distribusi kelas tidak seimbang. *Precision*, *recall*, dan *F-measure* digunakan untuk mengukur kinerja klasifikasi dari sisi ketepatan prediksi dan kemampuan model mengenali data pada suatu kelas [22], [23].

Balanced accuracy digunakan karena metrik ini didefinisikan sebagai rata-rata *recall* dari setiap kelas sehingga relevan untuk data tidak seimbang [24].

MCC digunakan sebagai metrik tambahan karena mempertimbangkan hubungan antara label aktual dan

prediksi model, serta mendukung evaluasi klasifikasi biner maupun *multikelas* [25], [26].

C adalah *confusion matrix* berukuran $K \times K$, dengan C_{ij} menyatakan jumlah data yang kelas aktualnya adalah kelas ke- i , tetapi diprediksi sebagai kelas ke- j . Jumlah kelas sentimen dinyatakan dengan K , sedangkan jumlah seluruh data uji dinyatakan dengan N . Berdasarkan notasi tersebut, *accuracy* dihitung menggunakan Persamaan (2).

$$Accuracy = \frac{\sum_{k=1}^K C_{kk}}{N} \quad (2)$$

Untuk kelas ke- k , nilai TP_k , FP_k , dan FN_k dihitung menggunakan Persamaan (3), Persamaan (4), dan Persamaan (5).

$$TP_k = C_{kk} \quad (3)$$

$$FP_k = \left(\sum_{i=1}^K C_{ik} \right) - C_{kk} \quad (4)$$

$$FN_k = \left(\sum_{j=1}^K C_{kj} \right) - C_{kk} \quad (5)$$

Precision, *recall*, dan *F1-score* pada kelas ke- k dihitung menggunakan Persamaan (6), Persamaan (7), dan Persamaan (8).

$$Precision_k = \frac{TP_k}{TP_k + FP_k} \quad (6)$$

$$Recall_k = \frac{TP_k}{TP_k + FN_k} \quad (7)$$

$$F1_k = 2 \times \frac{Precision_k \times Recall_k}{Precision_k + Recall_k} \quad (8)$$

Macro-F1 dihitung dengan merata-ratakan *F1-score* seluruh kelas tanpa mempertimbangkan jumlah data pada masing-masing kelas, sebagaimana ditunjukkan pada Persamaan (9).

$$Macro - F1 = \frac{1}{K} \sum_{k=1}^K F1_k \quad (9)$$

Weighted-F1 dihitung dengan mempertimbangkan proporsi jumlah data pada setiap kelas, sebagaimana ditunjukkan pada Persamaan (10).

$$Weighted - F1 = \sum_{k=1}^K \frac{N_k}{N} F1_k \quad (10)$$

Balanced accuracy dihitung sebagai rata-rata *recall* dari seluruh kelas, sebagaimana ditunjukkan pada Persamaan (11).

$$Balanced Accuracy = \frac{1}{K} \sum_{k=1}^K Recall_k \quad (11)$$

MCC untuk klasifikasi *multikelas* dihitung berdasarkan *confusion matrix*, sebagaimana ditunjukkan pada Persamaan (12).

$$MCC = \frac{c \times s - \sum_{k=1}^K p_k t_k}{\sqrt{(s^2 - \sum_{k=1}^K p_k^2)(s^2 - \sum_{k=1}^K t_k^2)}} \quad (12)$$

Pada Persamaan (2) sampai Persamaan (12), C_{ij} adalah elemen *confusion matrix* pada baris ke- i dan kolom ke- j , C_{kk} adalah jumlah prediksi benar pada kelas ke- k , TP_k adalah jumlah data kelas ke- k yang diprediksi benar, FP_k adalah jumlah data dari kelas lain yang salah diprediksi sebagai kelas ke- k , FN_k adalah jumlah data kelas ke- k yang salah diprediksi sebagai kelas lain, N_k adalah jumlah data aktual pada

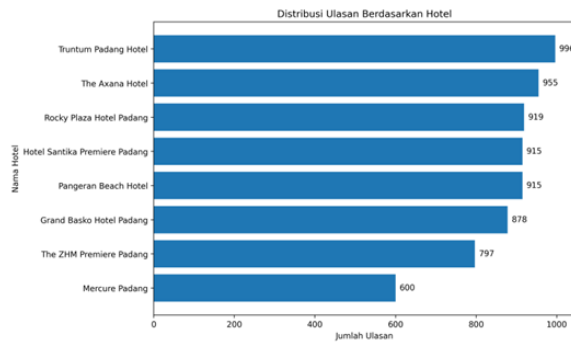
kelas ke- k , N adalah jumlah seluruh data uji, dan K adalah jumlah kelas sentimen. Pada Persamaan (12), $c = \sum_{k=1}^K C_{kk}$ adalah jumlah seluruh prediksi benar, $s = \sum_{i=1}^K \sum_{j=1}^K C_{ij}$ adalah jumlah seluruh data, $p_k = \sum_{i=1}^K C_{ik}$ adalah jumlah data yang diprediksi sebagai kelas ke- k , dan $t_k = \sum_{j=1}^K C_{kj}$ adalah jumlah data aktual pada kelas ke- k .

Training time dan *inference time* digunakan untuk mengevaluasi efisiensi komputasi. *Training time* dihitung sebagai selisih antara waktu selesai pelatihan dan waktu mulai pelatihan, sedangkan *inference time* dihitung sebagai selisih antara waktu selesai prediksi dan waktu mulai prediksi pada data uji. Selain metrik numerik, evaluasi juga dilakukan menggunakan *confusion matrix* dan *normalized confusion matrix* untuk melihat pola kesalahan prediksi antar kelas sentimen. *Classification report* digunakan untuk menyajikan *precision*, *recall*, dan *F1-score* pada setiap kelas. Analisis kesalahan prediksi dan analisis kata dominan pada setiap kelas digunakan sebagai bahan pendukung pembahasan hasil.

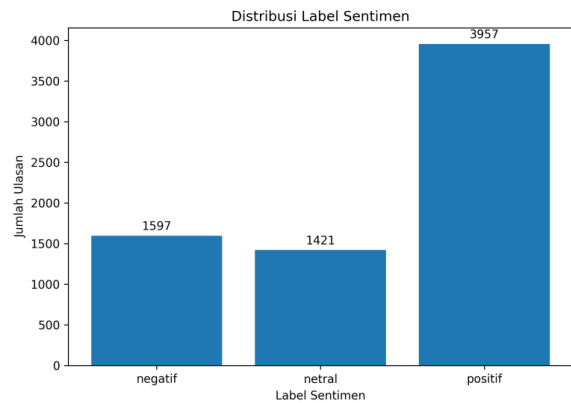
4. HASIL DAN PEMBAHASAN

4.1. Karakteristik Dataset

Dataset yang digunakan dalam penelitian ini terdiri dari 6.975 ulasan pelanggan hotel yang berasal dari platform Traveloka dan mencakup delapan hotel bintang empat di Kota Padang. Distribusi jumlah ulasan berdasarkan hotel menunjukkan bahwa setiap objek hotel memberikan kontribusi terhadap keseluruhan *dataset*. Sebagaimana ditunjukkan pada Gambar 3, jumlah ulasan per hotel berada pada rentang 600 hingga 996 ulasan, dengan kontribusi sekitar 8,60% hingga 14,28% dari total data. Pola tersebut menunjukkan bahwa data tidak terkonsentrasi secara berlebihan pada satu hotel tertentu, melainkan mencakup variasi ulasan dari beberapa hotel bintang empat di Kota Padang. Berdasarkan distribusi label sentimen pada Gambar 4, kelas positif merupakan kelas paling dominan dengan 3.957 ulasan atau 56,73% dari keseluruhan data, diikuti oleh kelas negatif sebanyak 1.597 ulasan atau 22,90% dan kelas netral sebanyak 1.421 ulasan atau 20,37%. Distribusi ini menunjukkan adanya ketidakseimbangan kelas, sehingga evaluasi model perlu mempertimbangkan metrik yang mampu menilai performa setiap kelas secara lebih proporsional. Selanjutnya, Gambar 5 menunjukkan bahwa distribusi sentimen pada setiap hotel tidak bersifat homogen. Perbedaan komposisi sentimen antar hotel mengindikasikan adanya variasi pola opini pelanggan pada masing-masing objek hotel. Dengan demikian, *dataset* ini memiliki sebaran ulasan antar hotel yang cukup bervariasi serta distribusi sentimen yang tidak seimbang, sehingga relevan digunakan untuk mengevaluasi kemampuan model dalam membedakan kelas negatif, netral, dan positif.



Gambar 3. Distribusi ulasan berdasarkan hotel

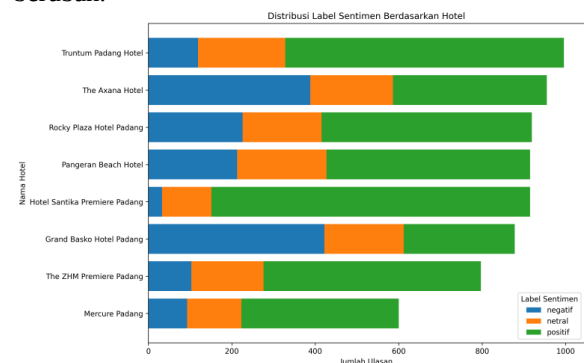


Gambar 4. Distribusi label sentimen

4.2. Hasil Pra-pemrosesan Data

Pra-pemrosesan data menghasilkan teks ulasan yang lebih seragam dan siap digunakan pada tahap

pemodelan. Hasil transformasi menunjukkan bahwa teks awal yang masih mengandung variasi penulisan, huruf kapital, tanda baca, *emoji*, istilah khusus hotel, kata tidak baku, dan pengulangan karakter dapat disederhanakan tanpa menghilangkan substansi opini pelanggan. Sebagaimana ditunjukkan pada Tabel 2, proses pra-pemrosesan mengubah bentuk seperti “tdk nyaman” menjadi “tidak nyaman”, *emoji* bermuatan sentimen menjadi token tekstual seperti *emoji_positif*, istilah “check-in” menjadi “checkin”, serta bentuk karakter berulang seperti “mantappp” menjadi bentuk yang lebih terkendali. Setelah proses pembersihan, teks yang kosong tidak digunakan pada tahap pemodelan, sedangkan kata-kata yang berpengaruh terhadap polaritas sentimen, terutama kata negasi seperti “tidak”, “kurang”, dan “belum”, tetap dipertahankan agar makna opini dalam ulasan tidak berubah.

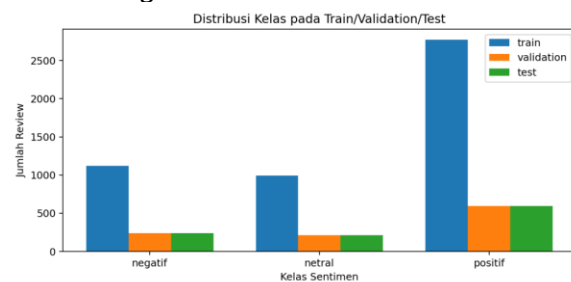


Gambar 5. Distribusi label sentimen berdasarkan hotel

Tabel 2. Contoh hasil pra-pemrosesan teks ulasan

Teks Ulasan Awal	Teks Setelah Pra-pemrosesan	Bentuk Transformasi
Rekomended pokoknya, anak-anak betah sampai enggak mau pulang.. ??	rekomendasi pokoknya anak anak betah sampai tidak mau pulang	Normalisasi kata tidak baku dan penghapusan karakter tidak relevan
Hotelnya ok banget, bersih dan rapi.	hotelnya oke banget bersih dan rapi	Normalisasi kata dan penghapusan tanda baca
Mulai check-in sampai check-out pelayanan sangat baik.	mulai checkin sampai checkout pelayanan sangat baik	Normalisasi istilah khusus
Staff hotel sangat ramah dan sangat membantu saat proses check-in.	staf hotel sangat ramah dan sangat membantu saat proses checkin	Normalisasi kata dan istilah khusus
mantapppppppppppp	mantapp	Pengurangan karakter berulang
Fasilitas seperti TV kurang, dan shower kamar mandi bermasalah	fasilitas seperti tv kurang dan shower kamar mandi bermasalah	Case folding dan penghapusan tanda baca

4.3. Pembagian Data



Gambar 6. Distribusi kelas sentimen pada data latih, validasi, dan uji

Dataset hasil pra-pemrosesan dibagi menjadi data latih, data validasi, dan data uji dengan proporsi 70:15:15. Hasil pembagian menunjukkan bahwa data latih terdiri dari 4.882 ulasan, data validasi terdiri dari 1.046 ulasan, dan data uji terdiri dari 1.047 ulasan. Sebagaimana ditunjukkan pada Gambar 6, distribusi kelas sentimen pada setiap *subset* tetap mengikuti pola distribusi *dataset* awal, yaitu kelas positif sebagai kelas dominan, diikuti oleh kelas negatif dan netral. Pada data latih, kelas negatif berjumlah 1.118 ulasan, kelas netral 995 ulasan, dan kelas positif 2.769 ulasan. Pada data validasi, kelas negatif berjumlah 239 ulasan, kelas netral 213 ulasan, dan kelas positif 594 ulasan, sedangkan pada data uji kelas negatif berjumlah 240

ulasan, kelas netral 213 ulasan, dan kelas positif 594 ulasan. Pola tersebut menunjukkan bahwa pembagian data secara *stratified* mampu mempertahankan proporsi kelas pada masing-masing *subset*, sehingga data validasi dan data uji tetap merepresentasikan distribusi sentimen pada *dataset* secara keseluruhan.

4.4. Hasil Evaluasi Kinerja Model

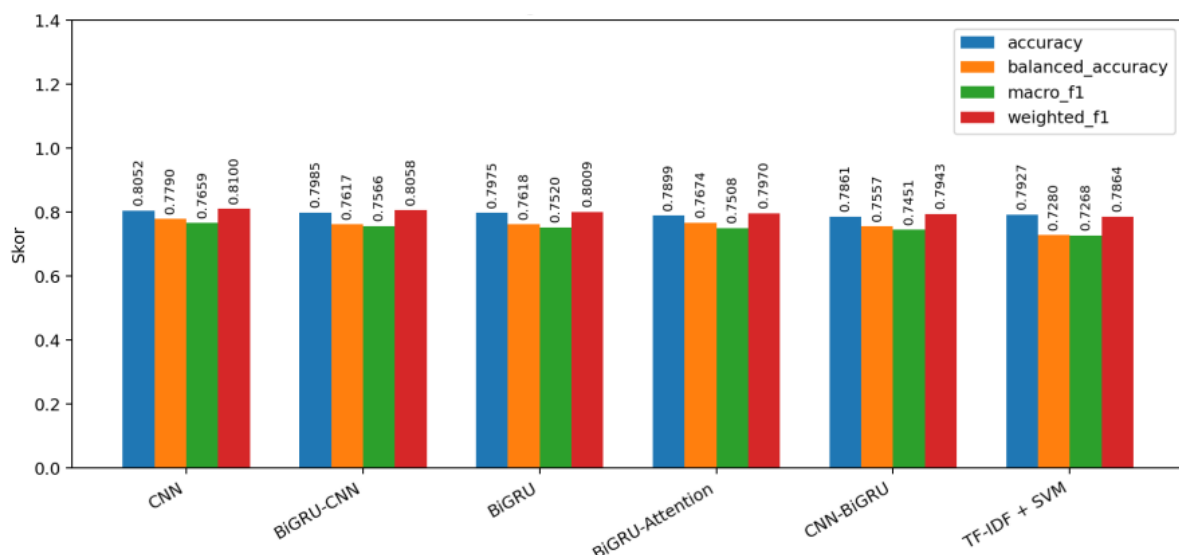
Hasil evaluasi pada data uji menunjukkan bahwa model berbasis jaringan saraf secara umum memperoleh nilai *macro-F1* lebih tinggi dibandingkan *TF-IDF + SVM* sebagai pembanding awal. Sebagaimana ditunjukkan pada Tabel 3, *CNN* memperoleh nilai *macro-F1* tertinggi sebesar 0.7659, dengan *accuracy* 0.8052, *balanced accuracy* 0.7790, *weighted-F1* 0.8100, dan *MCC* 0.6783. Berdasarkan *macro-F1* sebagai metrik utama, *CNN* menjadi model dengan kinerja paling baik dalam eksperimen ini. *BiGRU-CNN* menempati posisi berikutnya dengan *macro-F1* sebesar 0.7566, diikuti oleh *BiGRU* sebesar 0.7520, *BiGRU-Attention* sebesar 0.7508, *CNN-BiGRU* sebesar 0.7451, dan *TF-IDF + SVM* sebesar 0.7268. Perbandingan nilai metrik utama antar model juga divisualisasikan pada Gambar 7, yang memperlihatkan bahwa *CNN* cenderung unggul pada

beberapa metrik utama, terutama *accuracy*, *macro-F1*, dan *weighted-F1*.

Perbandingan antar arsitektur menunjukkan bahwa *TF-IDF + SVM* masih menghasilkan *accuracy* yang kompetitif, yaitu 0.7927, tetapi nilai *macro-F1* dan *balanced accuracy*-nya lebih rendah dibandingkan sebagian besar model berbasis jaringan saraf. Hal ini mengindikasikan bahwa *accuracy* saja belum cukup untuk menilai performa model pada *dataset* dengan distribusi kelas yang tidak seimbang. Pada kelompok model jaringan saraf, *CNN* sebagai model tunggal justru memperoleh *macro-F1* tertinggi dibandingkan model *hybrid* dan *attention-based*. Perbandingan antara *CNN-BiGRU* dan *BiGRU-CNN* menunjukkan bahwa urutan penggabungan komponen arsitektur memengaruhi kinerja model, karena *BiGRU-CNN* memperoleh *macro-F1* lebih tinggi dibandingkan *CNN-BiGRU*. Sementara itu, *BiGRU-Attention* belum menjadi model dengan *macro-F1* tertinggi, meskipun *balanced accuracy*-nya masih kompetitif dibandingkan beberapa model lain. Dengan demikian, mekanisme *attention* pada eksperimen ini belum memberikan peningkatan dominan dibandingkan *CNN* maupun *BiGRU-CNN*.

Tabel 3. Perbandingan kinerja model klasifikasi sentimen

Model	Accuracy	Balanced Accuracy	Macro Precision	Macro Recall	Macro-F1	Weighted-F1	MCC
CNN	0.8052	0.7790	0.7566	0.7790	0.7659	0.8100	0.6783
BiGRU-CNN	0.7985	0.7617	0.7568	0.7617	0.7566	0.8058	0.6646
BiGRU	0.7975	0.7618	0.7443	0.7618	0.7520	0.8009	0.6620
BiGRU-Attention	0.7899	0.7674	0.7410	0.7674	0.7508	0.7970	0.6581
CNN-BiGRU	0.7861	0.7557	0.7400	0.7557	0.7451	0.7943	0.6485
TF-IDF + SVM	0.7927	0.7280	0.7291	0.7280	0.7268	0.7864	0.6421



Gambar 7. Perbandingan metrik utama antar model

4.5. Analisis Efisiensi Komputasi

Analisis efisiensi komputasi dilakukan dengan membandingkan waktu pelatihan dan waktu inferensi pada setiap model. Sebagaimana ditunjukkan pada

Tabel 4, *TF-IDF + SVM* memiliki waktu pelatihan paling rendah, yaitu 0.3348 detik, dan waktu inferensi paling cepat, yaitu 0.0398 detik. Hal ini menunjukkan bahwa model pembanding awal berbasis *TF-IDF +*

SVM jauh lebih ringan secara komputasi dibandingkan model berbasis jaringan saraf. Pada kelompok model jaringan saraf, *CNN* memiliki waktu pelatihan paling rendah, yaitu 19.4247 detik, dan pada saat yang sama memperoleh *macro-F1* tertinggi pada evaluasi kinerja model. Sebaliknya, *BiGRU-CNN* memerlukan waktu pelatihan paling tinggi, yaitu 35.7335 detik, meskipun kinerjanya berada pada posisi kedua berdasarkan *macro-F1*. Temuan ini menunjukkan adanya *trade-off* antara kompleksitas arsitektur, performa klasifikasi, dan biaya komputasi. Model *hybrid* dan *recurrent* cenderung membutuhkan waktu pelatihan lebih lama dibandingkan *CNN*, sedangkan *TF-IDF* + *SVM* tetap menjadi model paling efisien dari sisi waktu pelatihan dan inferensi.

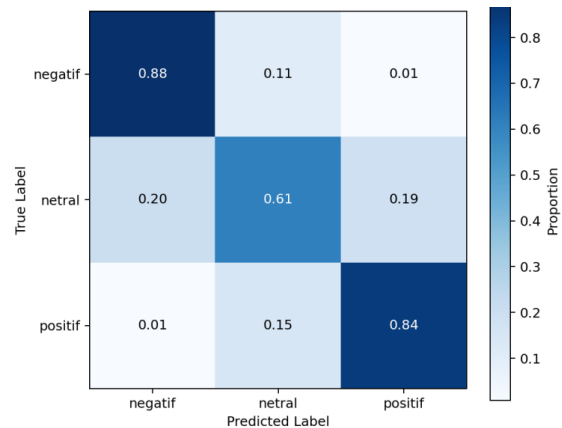
Tabel 4. perbandingan waktu pelatihan dan inferensi

Model	Training Time (s)	Inference Time (s)	Inference Time/Ulasan (ms)
TF-IDF + SVM	0.3348	0.0398	0.0380
CNN	19.4247	0.7540	0.7202
CNN-BiGRU	24.8496	0.6452	0.6162
BiGRU	28.2940	0.6194	0.5916
BiGRU-Attention	28.6062	0.7398	0.7066
BiGRU-CNN	35.7335	0.7675	0.7330

4.6. Analisis Confusion Matrix

Analisis *confusion matrix* dilakukan pada model *CNN* karena model tersebut memperoleh nilai *macro-F1* tertinggi pada evaluasi kinerja model. Sebagaimana ditunjukkan pada Gambar 8, hasil *normalized confusion matrix* memperlihatkan bahwa model mampu mengenali kelas negatif dan positif dengan proporsi prediksi benar yang lebih tinggi dibandingkan kelas netral. Kelas netral memiliki proporsi kesalahan yang lebih besar karena sebagian data netral diprediksi sebagai negatif dan sebagian lainnya diprediksi sebagai positif. Pola ini menunjukkan bahwa kelas netral merupakan kelas yang paling sulit dibedakan oleh model, terutama karena ulasan netral dapat memuat opini campuran, ekspresi yang tidak eksplisit,

atau informasi deskriptif tanpa kecenderungan sentimen yang kuat. Dengan demikian, hasil *confusion matrix* menunjukkan bahwa tantangan utama model bukan hanya membedakan ulasan negatif dan positif, tetapi juga mengidentifikasi batas antara kelas netral dan dua kelas polar lainnya.



Gambar 8. Normalized confusion matrix model CNN

4.7. Analisis Kesalahan Prediksi dan Kata Dominan

Analisis kesalahan prediksi dilakukan untuk mengidentifikasi pola ulasan yang masih sulit diklasifikasikan oleh model *CNN* sebagai model dengan nilai *macro-F1* tertinggi. Berdasarkan hasil *confusion matrix*, kesalahan klasifikasi terutama terjadi pada kelas netral, yaitu sebagian ulasan netral diprediksi sebagai negatif dan sebagian lainnya diprediksi sebagai positif. Pola ini mengindikasikan bahwa ulasan netral memiliki batas polaritas yang lebih sulit dikenali dibandingkan kelas negatif dan positif. Kesalahan juga dapat muncul pada ulasan positif yang memuat catatan minor atau ulasan negatif yang disampaikan secara tidak langsung, sehingga sinyal sentimen dalam teks menjadi kurang eksplisit. Ringkasan pola kesalahan prediksi model *CNN* disajikan pada Tabel 5.

Tabel 5. Ringkasan pola kesalahan prediksi model CNN

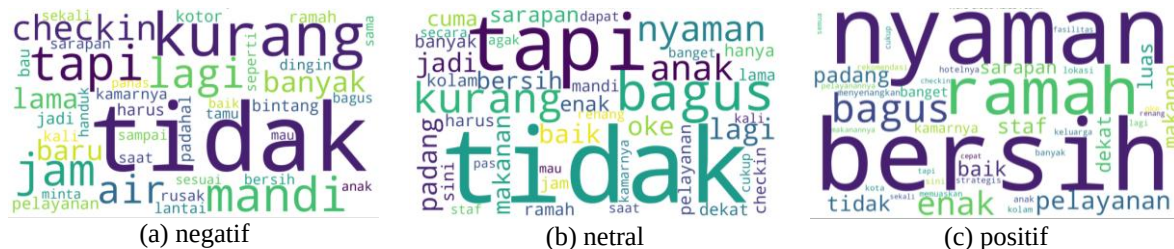
Pola Kesalahan	Karakteristik Ulasan	Kemungkinan Faktor Kesalahan
Netral → Negatif	Ulasan memuat informasi netral, tetapi disertai keluhan tertentu	Keluhan spesifik dapat menjadi sinyal negatif yang lebih kuat daripada keseluruhan isi ulasan
Netral → Positif	Ulasan bersifat informatif atau campuran, tetapi memuat kata bernuansa positif	Kata positif dapat mendominasi representasi teks meskipun label aktual netral
Positif → Netral	Ulasan cenderung positif, tetapi disertai catatan atau kekurangan minor	Adanya kata penghubung atau keluhan kecil dapat melemahkan polaritas positif
Negatif → Netral	Ulasan mengandung ketidakpuasan, tetapi disampaikan secara tidak langsung	Ekspresi negatif yang tidak eksplisit dapat membuat model gagal mengenali polaritas negatif
Positif/Negatif → Kelas berlawanan	Ulasan memuat opini campuran atau struktur kalimat kompleks	Kehadiran sinyal sentimen yang saling berlawanan dapat memengaruhi keputusan klasifikasi

Analisis kata dominan digunakan untuk melengkapi interpretasi kesalahan prediksi dengan melihat pola leksikal pada setiap kelas sentimen.

Visualisasi kata dominan disajikan dalam bentuk *word cloud* pada Gambar 11. Pada kelas negatif, kata-kata yang tampak menonjol berkaitan dengan keluhan atau

ketidakpuasan, seperti “tidak”, “kurang”, “mandi”, “air”, “jam”, dan “checkin”. Pada kelas netral, kata yang muncul cenderung bercampur antara kata bernuansa positif dan negatif, seperti “bagus”, “kurang”, “nyaman”, “baik”, dan “sarapan”, sehingga kelas ini cenderung berada di antara dua polaritas. Sementara itu, kelas positif didominasi oleh kata yang berkaitan dengan pengalaman menginap yang baik,

seperti “bersih”, “nyaman”, “ramah”, “bagus”, “pelayanan”, dan “sarapan”. Dengan demikian, analisis kesalahan prediksi dan *word cloud* menunjukkan bahwa tantangan utama model terletak pada ulasan yang ambigu, bercampur sentimen, atau memiliki batas polaritas yang tidak tegas, terutama pada kelas netral.



Gambar 11. *Word cloud* kata dominan pada setiap kelas sentimen: (a) negatif, (b) netral, dan (c) positif.

4.8. Pembahasan Komparatif

Hasil pengujian menunjukkan bahwa *CNN* memperoleh performa agregat tertinggi dibandingkan lima model lainnya. *CNN* menghasilkan *accuracy* sebesar 0.8052, *balanced accuracy* sebesar 0.7790, *macro-F1* sebesar 0.7659, *weighted-F1* sebesar 0.8100, dan *MCC* sebesar 0.6783. Berdasarkan *macro-F1* sebagai metrik utama, *CNN* mengungguli *BiGRU-CNN* sebesar 0.0093, *BiGRU* sebesar 0.0139, *BiGRU-Attention* sebesar 0.0151, *CNN-BiGRU* sebesar 0.0208, dan *TF-IDF + SVM* sebesar 0.0391. Meskipun *BiGRU-CNN* memperoleh *macro precision* sebesar 0.7568, sedikit lebih tinggi daripada *CNN* yang memperoleh 0.7566, *CNN* tetap unggul pada *accuracy*, *balanced accuracy*, *macro recall*, *macro-F1*, *weighted-F1*, dan *MCC*. Hasil tersebut menunjukkan bahwa keunggulan *CNN* tidak hanya didasarkan pada satu metrik, tetapi terlihat secara relatif konsisten pada sebagian besar metrik evaluasi.

Performa *CNN* yang lebih tinggi dapat diinterpretasikan berdasarkan kesesuaian mekanisme konvolusi dengan karakteristik teks ulasan hotel. *CNN* menggunakan filter konvolusi untuk mengekstraksi pola lokal dari beberapa token yang berdekatan. Pada penelitian ini, penggunaan ukuran kernel lima memungkinkan model mempelajari kombinasi kata atau frasa pendek yang berperan sebagai penanda sentimen. Ulasan hotel umumnya memuat ekspresi evaluatif yang relatif langsung, seperti “kamar sangat bersih”, “pelayanan kurang ramah”, “lokasi sangat strategis”, “air tidak lancar”, atau “fasilitas kurang lengkap”. Frasa-frasa tersebut dapat dikenali sebagai pola lokal tanpa selalu memerlukan pemodelan dependensi antarkata yang panjang. Kemampuan *CNN* dalam menangkap fitur lokal yang diskriminatif tersebut sesuai dengan karakteristik *CNN* untuk klasifikasi kalimat sebagaimana dijelaskan dalam penelitian Kim [5].

Interpretasi tersebut juga didukung oleh hasil analisis kata dominan. Kelas negatif memiliki kata-

kata yang berkaitan dengan keluhan, seperti “tidak”, “kurang”, “air”, “mandi”, dan “checkin”, sedangkan kelas positif memiliki kata-kata seperti “bersih” “nyaman”, “ramah”, “bagus”, dan “pelayanan”. Kemunculan penanda leksikal yang relatif eksplisit memberikan peluang bagi filter konvolusi untuk membentuk representasi fitur yang kuat. Dengan demikian, pada *dataset* penelitian ini, informasi sentimen kemungkinan lebih banyak direpresentasikan oleh kombinasi kata lokal daripada dependensi sekuensial yang panjang. Namun, penjelasan tersebut merupakan interpretasi berdasarkan karakteristik data dan hasil pengujian, sehingga tidak dapat diperlakukan sebagai bukti kausal tanpa analisis fitur atau visualisasi aktivasi filter *CNN* yang lebih lanjut.

Perbandingan antara *CNN* dan *TF-IDF + SVM* menunjukkan adanya *trade-off* antara performa prediktif dan efisiensi komputasi. *TF-IDF + SVM* menghasilkan *accuracy* sebesar 0.7927, yang relatif mendekati *accuracy CNN* sebesar 0.8052. Akan tetapi, *TF-IDF + SVM* hanya memperoleh *balanced accuracy* sebesar 0.7280, *macro-F1* sebesar 0.7268, *weighted-F1* sebesar 0.7864, dan *MCC* sebesar 0.6421. Dibandingkan *TF-IDF + SVM*, *CNN* meningkatkan *balanced accuracy* sebesar 0.0510, *macro-F1* sebesar 0.0391, *weighted-F1* sebesar 0.0236, dan *MCC* sebesar 0.0362. Selisih *balanced accuracy* dan *macro-F1* tersebut menunjukkan bahwa *CNN* memiliki kemampuan yang lebih seimbang dalam mengenali kelas negatif, netral, dan positif, sedangkan nilai *accuracy TF-IDF + SVM* masih dapat dipengaruhi oleh dominasi kelas positif dalam *dataset*.

Perbedaan tersebut dapat dikaitkan dengan representasi teks yang digunakan. *TF-IDF* merepresentasikan teks berdasarkan bobot kemunculan term dan tidak secara eksplisit mempertahankan urutan maupun hubungan semantik antarkata. Sebaliknya, *CNN* menggunakan representasi *embedding* dan operasi konvolusi untuk

mempelajari pola kombinasi kata yang berkontribusi terhadap sentimen. Meskipun demikian, *TF-IDF + SVM* tetap memiliki keunggulan yang sangat jelas dari sisi efisiensi. Model tersebut hanya memerlukan waktu pelatihan sebesar 0.3348 detik dan waktu inferensi sebesar 0.0398 detik, sedangkan *CNN* memerlukan waktu pelatihan sebesar 19.4247 detik dan waktu inferensi sebesar 0.7540 detik. Pada lingkungan eksperimen yang digunakan, *TF-IDF + SVM* dilatih sekitar 58 kali lebih cepat dan melakukan inferensi sekitar 18.9 kali lebih cepat daripada *CNN*. Oleh karena itu, *TF-IDF + SVM* tetap relevan untuk implementasi yang memprioritaskan keterbatasan sumber daya dan kecepatan pemrosesan, sedangkan *CNN* lebih sesuai apabila keseimbangan performa antar-kelas menjadi prioritas utama.

Dibandingkan dengan *BiGRU*, *CNN* memperoleh *macro-F1* yang lebih tinggi sebesar 0.0139, *balanced accuracy* lebih tinggi sebesar 0.0172, *weighted-F1* lebih tinggi sebesar 0.0091, dan *MCC* lebih tinggi sebesar 0.0163. *BiGRU* dirancang untuk memproses konteks sekuensial dari arah maju dan mundur sehingga secara konseptual mampu mempertimbangkan konteks sebelum dan sesudah suatu kata [6], [13].

Namun, kemampuan tersebut belum menghasilkan performa yang melampaui *CNN* pada *dataset* penelitian ini. Salah satu penjelasan yang mungkin adalah bahwa banyak ulasan hotel mengandung penanda sentimen yang bersifat lokal dan eksplisit, sehingga tambahan pemodelan sekuensial dua arah belum memberikan informasi diskriminatif yang lebih besar. *BiGRU* juga memerlukan waktu pelatihan sebesar 28.2940 detik, sedangkan *CNN* hanya memerlukan 19.4247 detik. Dengan demikian, *CNN* memperoleh *macro-F1* lebih tinggi dengan waktu pelatihan sekitar 31.35% lebih singkat daripada *BiGRU*.

Hasil model hibrida menunjukkan bahwa urutan penggabungan *CNN* dan *BiGRU* memengaruhi performa klasifikasi. *BiGRU-CNN* memperoleh *macro-F1* sebesar 0.7566, sedangkan *CNN-BiGRU* memperoleh 0.7451. Dengan demikian, *BiGRU-CNN* lebih tinggi sebesar 0.0115 dibandingkan *CNN-BiGRU*. Pada *BiGRU-CNN*, informasi sekuensial dua arah dibentuk terlebih dahulu, kemudian *CNN* mengekstraksi pola lokal dari representasi sekuensial tersebut. Urutan ini kemungkinan menghasilkan representasi yang lebih diskriminatif dibandingkan *CNN-BiGRU*, yang mengekstraksi fitur lokal terlebih dahulu sebelum memprosesnya secara sekuensial. Meskipun demikian, kedua arsitektur hibrida tersebut belum melampaui *CNN* tunggal. *BiGRU-CNN* masih berada 0.0093 di bawah *CNN* berdasarkan *macro-F1*, sedangkan *CNN-BiGRU* berada 0.0208 di bawah *CNN*.

Hasil tersebut menunjukkan bahwa penambahan komponen arsitektur tidak secara otomatis meningkatkan performa klasifikasi. Pemrosesan berlapis pada model hibrida dapat meningkatkan

kapasitas representasi, tetapi pada ukuran dan karakteristik *dataset* tertentu juga dapat menambah kompleksitas optimasi dan jumlah parameter yang harus dipelajari. *BiGRU-CNN* memerlukan waktu pelatihan sebesar 35.7335 detik atau sekitar 45.64% lebih lama dibandingkan *CNN*, sedangkan *CNN-BiGRU* memerlukan waktu pelatihan sebesar 24.8496 detik atau sekitar 21.83% lebih lama dibandingkan *CNN*. Oleh karena itu, peningkatan kompleksitas kedua model hibrida belum menghasilkan peningkatan performa yang sebanding dengan tambahan biaya komputasinya. Temuan ini menunjukkan bahwa model yang lebih kompleks belum tentu lebih sesuai apabila karakteristik tugas dapat dimodelkan secara efektif oleh arsitektur yang lebih sederhana.

BiGRU-Attention memperoleh *balanced accuracy* sebesar 0.7674, yang merupakan nilai kedua tertinggi setelah *CNN*. Hasil tersebut menunjukkan bahwa mekanisme *attention* masih memiliki kemampuan yang kompetitif dalam menjaga pengenalan antar-kelas. Akan tetapi, *BiGRU-Attention* memperoleh *macro precision* sebesar 0.7410 dan *macro-F1* sebesar 0.7508, lebih rendah daripada *CNN*. Mekanisme *attention* secara konseptual memberikan bobot lebih besar kepada bagian teks yang dianggap relevan [14].

Namun, pada ulasan netral atau ulasan dengan opini campuran, kata positif dan negatif dapat muncul dalam satu teks yang sama. Kondisi tersebut dapat menyebabkan bagian-bagian yang memperoleh bobot tinggi tidak selalu mengarah pada satu polaritas yang konsisten. Selain itu, keuntungan mekanisme *attention* dapat dipengaruhi oleh jumlah data, panjang teks, proses optimasi, dan konfigurasi *hiperparameter*. Oleh karena bobot *attention* belum dianalisis secara langsung dalam penelitian ini, penjelasan tersebut harus dipahami sebagai interpretasi hasil, bukan sebagai kesimpulan kausal.

Dari sisi efisiensi model jaringan saraf, *CNN* memiliki waktu pelatihan terendah, yaitu 19.4247 detik. Namun, *CNN* tidak memiliki waktu inferensi tercepat di antara model neural. Waktu inferensi *CNN* sebesar 0.7540 detik masih lebih tinggi daripada *BiGRU* sebesar 0.6194 detik, *CNN-BiGRU* sebesar 0.6452 detik, dan *BiGRU-Attention* sebesar 0.7398 detik. Oleh karena itu, keunggulan efisiensi *CNN* dalam penelitian ini terutama terlihat pada waktu pelatihan, bukan pada seluruh aspek komputasi. Meskipun demikian, perbedaan waktu inferensi per ulasan relatif kecil, sedangkan *CNN* tetap memperoleh performa prediktif tertinggi. Berdasarkan kondisi tersebut, *CNN* memberikan kompromi yang paling baik antara performa klasifikasi dan kompleksitas pelatihan pada eksperimen ini.

Analisis *confusion matrix* menunjukkan bahwa kelas netral merupakan kelas yang paling sulit dikenali. Sebagian ulasan netral diprediksi sebagai negatif dan sebagian lainnya diprediksi sebagai positif. Kesalahan tersebut dapat terjadi karena label netral mencakup ulasan informatif, ulasan ambigu, dan

ulasan dengan opini campuran yang tidak memiliki polaritas dominan. Sebagai contoh, ulasan dapat memberikan penilaian positif terhadap kebersihan tetapi pada saat yang sama memuat keluhan terhadap fasilitas atau pelayanan. Dalam kondisi tersebut, kata bermuatan positif atau negatif dapat mendominasi representasi model dan menggeser prediksi dari kelas netral ke salah satu kelas polar.

Kesulitan mengenali kelas netral juga menjelaskan pentingnya penggunaan *macro-F1*, *balanced accuracy*, dan *MCC* sebagai pelengkap *accuracy*. *Dataset* penelitian didominasi kelas positif sebanyak 3.957 ulasan atau 56.73%, sedangkan kelas negatif berjumlah 1.597 ulasan atau 22.90% dan kelas netral berjumlah 1.421 ulasan atau 20.37%. Pada distribusi tersebut, *accuracy* yang tinggi belum tentu menunjukkan bahwa model mampu mengenali semua kelas secara seimbang. *Macro-F1* memberikan bobot yang sama kepada setiap kelas, *balanced accuracy* menghitung rata-rata *recall antarkelas*, sedangkan *MCC* mempertimbangkan kesesuaian keseluruhan antara label aktual dan prediksi. Keunggulan *CNN* pada ketiga metrik tersebut memperkuat kesimpulan bahwa *CNN* tidak hanya menguntungkan kelas mayoritas, tetapi memiliki keseimbangan klasifikasi yang lebih baik dibandingkan model lain pada pengujian ini.

Temuan penelitian ini memiliki keterkaitan dengan penelitian Kusumaningrum et al. [3], yang menunjukkan bahwa *CNN* dapat memberikan performa kompetitif dibandingkan metode pembelajaran mesin klasik pada ulasan hotel Indonesia. Penelitian ini memperluas perbandingan tersebut dengan melibatkan *BiGRU*, dua susunan model hibrida, dan *BiGRU-Attention* dalam klasifikasi tiga kelas. Berbeda dengan asumsi bahwa peningkatan kompleksitas arsitektur selalu meningkatkan performa, hasil penelitian menunjukkan bahwa *CNN* tunggal justru memperoleh *macro-F1* tertinggi. Hasil tersebut menegaskan bahwa kesesuaian arsitektur dengan karakteristik data lebih penting daripada kompleksitas model semata. Pada data yang banyak mengandung frasa evaluatif lokal, *CNN* dapat menghasilkan representasi yang memadai tanpa memerlukan tambahan komponen *recurrent* atau *attention*.

Secara praktis, *CNN* dapat dipertimbangkan sebagai model utama untuk klasifikasi sentimen ulasan hotel dengan karakteristik data yang serupa, khususnya ketika tujuan implementasi menekankan keseimbangan pengenalan kelas dan waktu pelatihan yang relatif rendah. *TF-IDF + SVM* dapat dipilih ketika efisiensi komputasi menjadi pertimbangan utama, sedangkan *BiGRU*, model hibrida, dan *BiGRU-Attention* dapat dipertimbangkan apabila data memiliki dependensi sekuensial yang lebih kompleks atau tersedia *dataset* yang lebih besar. Dengan demikian, pemilihan model sebaiknya tidak hanya didasarkan pada nilai *accuracy*, tetapi juga mempertimbangkan *macro-F1*, *balanced accuracy*,

MCC, waktu komputasi, karakteristik bahasa, dan kebutuhan implementasi.

Meskipun *CNN* memperoleh nilai tertinggi pada penelitian ini, hasil tersebut perlu ditafsirkan sesuai batas eksperimen. *Dataset* hanya berasal dari ulasan Traveloka pada hotel bintang empat di Kota Padang, sedangkan pengujian dilakukan menggunakan satu pembagian data 70:15:15 dan satu *random seed*, yaitu 42. Penelitian ini juga belum melaporkan variasi performa dari beberapa pengulangan eksperimen maupun uji signifikansi statistik terhadap perbedaan antar-model. Oleh karena itu, selisih performa, khususnya antara *CNN* dan *BiGRU-CNN* yang hanya sebesar 0.0093 pada *macro-F1*, belum dapat dinyatakan signifikan secara statistik. Pengujian dengan beberapa *random seed*, *repeated stratified split*, interval kepercayaan, atau uji signifikansi diperlukan untuk memastikan kestabilan peringkat model. Generalisasi hasil juga perlu diuji pada platform OTA, kota, kategori hotel, dan ukuran *dataset* yang berbeda.

5. KESIMPULAN DAN SARAN

Berdasarkan hasil pengujian terhadap 1.047 data uji, *CNN* memperoleh performa tertinggi di antara enam model klasifikasi sentimen yang dievaluasi. *CNN* menghasilkan *accuracy* sebesar 0,8052, *balanced accuracy* sebesar 0,7790, *macro-F1* sebesar 0,7659, *weighted-F1* sebesar 0,8100, dan *MCC* sebesar 0,6783. Berdasarkan *macro-F1* sebagai metrik utama, *CNN* mengungguli *BiGRU-CNN* yang memperoleh 0,7566, *BiGRU* sebesar 0,7520, *BiGRU-Attention* sebesar 0,7508, *CNN-BiGRU* sebesar 0,7451, dan *TF-IDF + SVM* sebesar 0,7268. *CNN* juga memiliki waktu pelatihan paling rendah di antara model jaringan saraf, yaitu 19,4247 detik, dengan waktu inferensi sebesar 0,7540 detik. Sementara itu, *TF-IDF + SVM* merupakan model paling efisien secara komputasi dengan waktu pelatihan 0,3348 detik dan waktu inferensi 0,0398 detik, tetapi memiliki *balanced accuracy* dan *macro-F1* yang lebih rendah dibandingkan model jaringan saraf. Hasil tersebut menunjukkan bahwa *CNN* memberikan keseimbangan terbaik antara performa klasifikasi dan biaya pelatihan pada *dataset* penelitian ini. Analisis *confusion matrix* juga menunjukkan bahwa kelas netral menjadi kelas yang paling sulit dikenali karena ulasannya cenderung ambigu, bersifat informatif, atau memuat opini positif dan negatif secara bersamaan. Dengan demikian, keunggulan *CNN* dalam penelitian ini terutama berkaitan dengan kemampuannya menangkap pola lokal dan frasa evaluatif yang menjadi penanda sentimen dalam ulasan hotel.

Berdasarkan hasil tersebut, penelitian selanjutnya disarankan memprioritaskan peningkatan kemampuan model dalam mengenali kelas netral. Upaya tersebut dapat dilakukan melalui penyempurnaan pedoman anotasi, pengukuran kesepakatan antar-anotator, penambahan data kelas negatif dan netral, serta penerapan teknik penanganan ketidakseimbangan

kelas. Analisis sentimen berbasis aspek juga dapat digunakan untuk memisahkan sentimen terhadap kamar, fasilitas, pelayanan, lokasi, kebersihan, dan harga dalam satu ulasan sehingga opini campuran tidak langsung dipaksakan ke dalam satu label sentimen. Selain itu, kestabilan performa CNN dan model lainnya perlu diuji melalui beberapa *random seed*, pengulangan eksperimen, interval kepercayaan, dan uji signifikansi statistik karena penelitian ini masih menggunakan satu pembagian data dengan *random seed* 42. Pengujian pada platform OTA, kota, kategori hotel, dan jumlah data yang berbeda juga diperlukan untuk mengetahui kemampuan generalisasi model. Model berbasis *transformer* dapat dibandingkan pada penelitian berikutnya, tetapi evaluasinya tetap perlu mempertimbangkan peningkatan *macro-F1*, *balanced accuracy*, *MCC*, dan biaya komputasi agar penggunaan model yang lebih kompleks didukung oleh peningkatan performa yang sebanding.

DAFTAR PUSTAKA

- [1] A. Ameer, S. Hamdi, and S. ben Yahia, "Sentiment Analysis for Hotel Reviews: A Systematic Literature Review," *ACM Computing Surveys*, vol. 56, no. 2, pp. 1–38, Feb. 2024, doi: 10.1145/3605152.
- [2] R. Kusumaningrum, I. Z. Nisa, R. Jayanto, R. P. Nawangsari, and A. Wibowo, "Deep learning-based application for multilevel sentiment analysis of Indonesian hotel reviews," *Heliyon*, vol. 9, no. 6, p. e17147, Jun. 2023, doi: 10.1016/j.heliyon.2023.e17147.
- [3] R. Kusumaningrum, I. Z. Nisa, R. P. Nawangsari, and A. Wibowo, "Sentiment analysis of Indonesian hotel reviews: from classical machine learning to deep learning," *International Journal of Advances in Intelligent Informatics*, vol. 7, no. 3, p. 292, Nov. 2021, doi: 10.26555/ijain.v7i3.737.
- [4] V. W. D. Thomas and F. Rumaisa, "Analisis Sentimen Ulasan Hotel Bahasa Indonesia Menggunakan Support Vector Machine dan TF-IDF," *JURNAL MEDIA INFORMATIKA BUDIDARMA*, vol. 6, no. 3, p. 1767, Jul. 2022, doi: 10.30865/mib.v6i3.4218.
- [5] Y. Kim, "Convolutional Neural Networks for Sentence Classification," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2014, pp. 1746–1751. doi: 10.3115/v1/D14-1181.
- [6] K. Cho *et al.*, "Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2014, pp. 1724–1734. doi: 10.3115/v1/D14-1179.
- [7] Z. Gao, Z. Li, J. Luo, and X. Li, "Short Text Aspect-Based Sentiment Analysis Based on CNN + BiGRU," *Applied Sciences*, vol. 12, no. 5, p. 2707, Mar. 2022, doi: 10.3390/app12052707.
- [8] J. Zhu and H. Li, "Text Sentiment Analysis Based on CNN-BiGRU Enhanced Features," in *Proceedings of the 2022 10th International Conference on Information Technology: IoT and Smart City*, New York, NY, USA: ACM, Dec. 2022, pp. 15–21. doi: 10.1145/3582197.3582200.
- [9] M. M. Mathew, P. Nanjundan, and S. Bashir, "Sentiment Analysis of Online Hotel Reviews Employing Bidirectional GRU with Attention Mechanism," in *2024 International Conference on Emerging Systems and Intelligent Computing (ESIC)*, IEEE, Feb. 2024, pp. 237–242. doi: 10.1109/ESIC60604.2024.10481617.
- [10] S. Cahyaningtyas, D. Hatta Fudholi, and A. Fathan Hidayatullah, "Deep Learning for Aspect-Based Sentiment Analysis on Indonesian Hotels Reviews," *Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control*, Aug. 2021, doi: 10.22219/kinetik.v6i3.1300.
- [11] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Information Processing & Management*, vol. 24, no. 5, pp. 513–523, Jan. 1988, doi: 10.1016/0306-4573(88)90021-0.
- [12] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, Sep. 1995, doi: 10.1007/BF00994018.
- [13] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997, doi: 10.1109/78.650093.
- [14] D. Bahdanau, K. Cho, and Y. Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate," *arXiv:1409.0473*, May 2016. doi: 10.48550/arXiv.1409.0473.
- [15] H. Çetiner and S. Metlek, "SAHRAN: Sentiment Analysis of Hotel Reviews with Attention-Based Recurrent Neural Network," *Iğdır Üniversitesi Fen Bilimleri Enstitüsü Dergisi*, vol. 15, no. 1, pp. 39–56, Mar. 2025, doi: 10.21597/jist.1523220.
- [16] C. S. R. Priya and P. Deepalakshmi, "Sentiment analysis from unstructured hotel reviews data in social network using deep learning techniques," *International Journal of Information Technology*, vol. 15, no. 7, pp. 3563–3574, Oct. 2023, doi: 10.1007/s41870-023-01419-z.
- [17] D. Erdoğan *et al.*, "Developing a Deep Learning-Based Sentiment Analysis System of Hotel Customer Reviews for Sustainable Tourism," *Sustainability*, vol. 17, no. 13, p. 5756, Jun. 2025, doi: 10.3390/su17135756.

- [18] B. Yanuargi, E. Utami, K. Kusrini, and A. A. Parikesit, "Improving Model Capability for Sentiment Trend Analysis in Hotel Visitor Reviews with Bi-LSTM Multistage Approach," *Jurnal Teknik Informatika (Jutif)*, vol. 6, no. 5, pp. 3652–3666, Oct. 2025, doi: 10.52436/1.jutif.2025.6.5.5185.
- [19] W. Chen, K. Yang, Z. Yu, Y. Shi, and C. L. P. Chen, "A survey on imbalanced learning: latest research, applications and future directions," *Artificial Intelligence Review*, vol. 57, no. 6, p. 137, May 2024, doi: 10.1007/s10462-024-10759-6.
- [20] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *arXiv:1412.6980*, Jan. 2017, doi: 10.48550/arXiv.1412.6980.
- [21] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A Simple Way to Prevent Neural Networks from Overfitting," *Journal of Machine Learning Research*, vol. 15, pp. 1929–1958, 2014.
- [22] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information Processing & Management*, vol. 45, no. 4, pp. 427–437, Jul. 2009, doi: 10.1016/j.ipm.2009.03.002.
- [23] D. M. W. Powers, "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation," *arXiv:2010.16061*, Oct. 2020, doi:10.48550/arXiv.2010.16061.
- [24] K. H. Brodersen, C. S. Ong, K. E. Stephan, and J. M. Buhmann, "The Balanced Accuracy and Its Posterior Distribution," in *2010 20th International Conference on Pattern Recognition*, IEEE, Aug. 2010, pp. 3121–3124, doi: 10.1109/ICPR.2010.764.
- [25] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, no. 1, p. 6, Dec. 2020, doi: 10.1186/s12864-019-6413-7.
- [26] J. Gorodkin, "Comparing two K-category assignments by a K-category correlation coefficient," *Computational Biology and Chemistry*, vol. 28, no. 5–6, pp. 367–374, Dec. 2004, doi: 10.1016/j.compbiolchem.2004.09.006.