Jurnal Dialektika Informatika (Detika) DOI: ...

Vol. x, No. x, Mei 2026, hlm. x-x p-ISSN: 2746-2811
e-ISSN: xxxx-xxxx

IMPLEMENTASI *NAÏVE BAYES* DALAM PENGOLAHAN SENTIMEN PENGGUNA *TIKTOK* TERHADAP HOTEL DI KOTA SINGKAWANG

Hikmal[1], Fathushahib[2], Heldi Hastriyandi[3], Sri Wahyuni[4]

[1]Program Studi D3 Manajemen Informatika, Politeknik Negeri Sambas
Jl. Raya Sejangkung, Sambas, Kalimantan Barat 79462
Email: [1]hikmalhikmal346@gmail.com

(Naskah masuk: ___/___/2026, diterima untuk diterbitkan: ___/___/2026)

Abstrak

Perkembangan pariwisata di Kota Singkawang mendorong wisatawan membagikan pengalaman menginap melalui komentar *TikTok* yang bersifat spontan, namun volumenya yang besar menyulitkan analisis manual. Penelitian ini menerapkan *Multinomial Naïve Bayes* dengan pembobotan *TF-IDF* untuk mengklasifikasikan sentimen komentar *TikTok* terhadap empat hotel di Kota Singkawang, menggunakan pendekatan *Hybrid Lexicon* yang mengombinasikan *InSet Lexicon* dengan filter kata ambigu, filter kata topikal netral, penambahan leksikon domain hotel (±90 kata), penanganan negasi berbasis *window*, serta ambang bukti minimum ($|skor| \geq 2$) untuk mengurangi kesalahan pelabelan pada kalimat bermuatan lemah. Dari 2.266 komentar mentah yang dihimpun dari 32 file CSV, diperoleh 1.758 komentar bersih (reduksi 22,4%) setelah tujuh tahap preprocessing. Hasil pelabelan menunjukkan distribusi Netral mendominasi sebesar 57,8% (1.017 komentar), diikuti Positif 28,6% (502) dan Negatif 13,6% (239), dengan rasio Positif:Negatif 2,10:1. Model dilatih dengan skema pembagian data 80:20 secara *stratified* dan menghasilkan rata-rata akurasi 73,02%, *precision* 0,68, *recall* 0,73, dan *F1-score* 0,67, unggul 39,7 poin persentase di atas baseline random classifier tiga kelas (33,3%). Analisis *TF-IDF* menunjukkan fitur khas tiap hotel, seperti "mantap" pada Hotel Mahkota dan "pantai" pada Hotel Horison Ultima, yang dapat menjadi masukan evaluatif bagi manajemen hotel.

Kata kunci: *analisis sentimen, multinomial naïve bayes, Hybrid Lexicon, TF-IDF, TikTok*

IMPLEMENTATION OF NAÏVE BAYES IN PROCESSING TIKTOK USER SENTIMENT TOWARD HOTELS IN SINGKAWANG CITY

Abstract

The growth of tourism in Singkawang City encourages tourists to share their lodging experiences through spontaneous TikTok comments, yet the large volume makes manual analysis impractical. This study applies Multinomial Naïve Bayes with TF-IDF weighting to classify the sentiment of TikTok comments toward four hotels in Singkawang City using a Hybrid Lexicon approach that combines the InSet Lexicon with an ambiguous-word filter, a neutral-topical-word filter, an additional hotel-domain lexicon (±90 words), window-based negation handling, and a minimum evidence threshold ($|score| \geq 2$) to reduce mislabeling of weakly polarized sentences. From 2,266 raw comments collected from 32 CSV files, 1,758 clean comments (22.4% reduction) were obtained after seven preprocessing stages. The labeling results show that the Neutral class dominates at 57.8% (1,017 comments), followed by Positive at 28.6% (502) and Negative at 13.6% (239), with a Positive:Negative ratio of 2.10:1. The model was trained using an 80:20 stratified split and achieved an average accuracy of 73.02%, precision of 0.68, recall of 0.73, and F1-score of 0.67, outperforming the three-class random-classifier baseline (33.3%) by 39.7 percentage points. TF-IDF analysis revealed hotel-specific distinctive features, such as "mantap" for Hotel Mahkota and "pantai" for Hotel Horison Ultima, which can serve as evaluative input for hotel management.

Keywords: *sentiment analysis, multinomial naïve bayes, Hybrid Lexicon, TF-IDF, Tiktok*

Vol. x, No. x, Mei 2026, hlm. x-x p-ISSN: 2746-2811
e-ISSN: xxxx-xxxx

1. PENDAHULUAN

Perkembangan sektor pariwisata di Kota Singkawang mendorong persaingan antar hotel dalam memberikan pelayanan terbaik kepada wisatawan. *TikTok* kini menjadi salah satu media utama yang digunakan wisatawan untuk membagikan pengalaman menginap mereka melalui komentar yang bersifat spontan dan real-time, sehingga menjadi sumber data yang kaya namun belum banyak dimanfaatkan secara sistematis oleh pelaku industri perhotelan setempat.

Banyaknya jumlah komentar tersebut membuat proses analisis secara manual menjadi kurang efektif dan efisien, sehingga diperlukan pendekatan analisis sentimen berbasis *machine learning* untuk mengelompokkan opini pengguna ke dalam kategori Positif, Negatif, dan Netral secara otomatis (Liu, 2020). Namun, pelabelan sentimen berbasis leksikon murni memiliki keterbatasan: kata-kata ambigu, istilah topikal netral, dan kosakata khas domain hotel yang tidak terdaftar pada kamus umum berpotensi menghasilkan label yang keliru, terutama pada kalimat dengan muatan sentimen yang lemah.

Oleh karena itu, penelitian ini menerapkan algoritma *Multinomial Naïve Bayes* dengan pembobotan TF-IDF, dilengkapi pendekatan *Hybrid Lexicon* yang mengombinasikan *InSet Lexicon* dengan filter kata ambigu, filter kata topikal netral, penambahan leksikon domain hotel, penanganan negasi berbasis *window*, serta ambang bukti minimum, untuk menganalisis komentar *TikTok* pada empat hotel di Kota Singkawang: Hotel Mahkota Singkawang, Swiss-Belhotel Singkawang, Dayang Resort Singkawang, dan Horison Ultima Singkawang.

Rumusan masalah penelitian ini adalah: (1) bagaimana distribusi sentimen komentar pengguna *TikTok* terhadap empat hotel di Kota Singkawang berdasarkan pendekatan *Hybrid Lexicon*, dan (2) bagaimana performa algoritma *Multinomial Naïve Bayes* dengan pembobotan *TF-IDF* dalam mengklasifikasikan sentimen tersebut. Tujuan penelitian ini sejalan dengan kedua rumusan masalah, yaitu mengetahui distribusi sentimen serta mengukur performa dan tingkat akurasi model yang dibangun.

2. TINJAUAN PUSTAKA

2.1. Analisis Sentimen

Analisis sentimen merupakan cabang *Natural Language Processing* (NLP) yang bertujuan mengidentifikasi dan mengklasifikasikan orientasi opini dalam teks ke dalam kategori Positif, Negatif, atau Netral (Medhat, Hassan, & Korashy, 2014). Dalam konteks industri perhotelan, analisis sentimen dapat dimanfaatkan untuk memahami persepsi tamu secara cepat tanpa membaca seluruh ulasan secara manual.

2.2. *Multinomial Naïve Bayes*

Naïve Bayes adalah algoritma klasifikasi probabilistik berbasis *Teorema Bayes* dengan asumsi independensi antar fitur [4]. Formula dasar: $P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$. Dalam konteks klasifikasi teks:

$$P(c \mid d) \propto P(c) \prod_{i=1}^{n} P(w_i \mid c)$$

Varian *Multinomial* dipilih karena sesuai untuk data frekuensi kata seperti yang dihasilkan *TF-IDF*. *Laplace Smoothing (alpha=1,0)* diterapkan untuk mencegah *zero probability* pada kata yang tidak muncul di data latih.

2.3. *InSet Lexicon* dan *Hybrid Lexicon*

InSet Lexicon merupakan kamus sentimen berbahasa Indonesia yang dikembangkan oleh Koto dan Mahendra (2017), berisi kata-kata yang telah diberi skor polaritas positif maupun negatif. Pendekatan leksikon murni memiliki kelemahan berupa tingginya sensitivitas terhadap kata ambigu dan kata topikal netral. Penelitian ini mengembangkan pendekatan *Hybrid Lexicon* yang menggabungkan *InSet Lexicon* dengan empat komponen tambahan: (1) filter kata ambigu, (2) filter kata topikal netral, (3) leksikon tambahan domain hotel sebanyak kurang lebih 90 kata, dan (4) ambang bukti minimum $|\text{skor}| \geq 2$ pada tahap agregasi skor kalimat, sehingga kalimat dengan bukti sentimen yang terlalu lemah akan diklasifikasikan sebagai Netral alih-alih dipaksakan ke kelas Positif atau Negatif.

2.4. TF-IDF (*Term Frequency–Inverse Document Frequency)*

TF-IDF merepresentasikan teks sebagai vektor numerik: $TF\text{-}IDF(t,d) = TF(t,d) \times IDF(t)$, di mana $IDF(t) = \log\left(\frac{N}{df(t)}\right)$ [17]. Parameter yang digunakan*:max_features=1.000,ngram_range=(1,2)* untuk menangkap *unigram* dan *bigram, sublinear_tf=True*. Teknik *anti-leakage* diterapkan: *TF-IDF di-fit* eksklusif pada data latih menggunakan *fit_transform(X_train)* dan hanya ditransformasi pada data uji menggunakan *transform(X_test)* tanpa fit ulang, untuk mencegah kebocoran informasi data uji ke proses pembelajaran.

2.5. Penelitian Terdahulu

Beberapa penelitian terdahulu yang relevan dirangkum pada Tabel 1.

Tabel 1. Penelitian Terdahulu

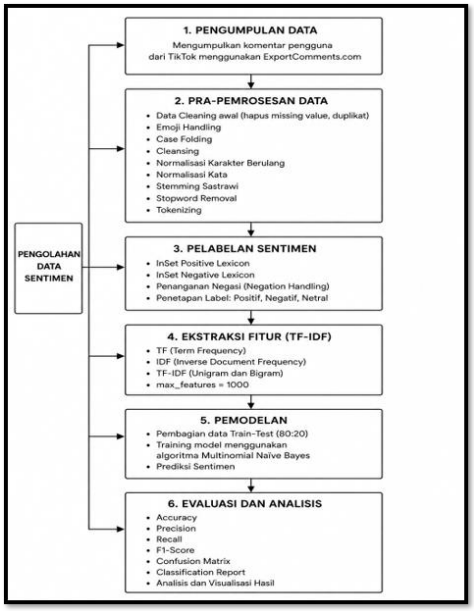
Penulis (Tahun)	Platform	Akurasi	Perbedaan

Azzahra dkk. (2024)	*Google Play*	F1 baik	Ulasan app *TikTok*
Thomas & Rumaisa (2022)	*Tripadvisor*	85%	SVM, platform formal
Rifa'i dkk. (2021)	*Google Maps*	76%	Wisata Kalbar umum
Azlan & Rahmi (2025)	*Google Review*	Kompetitif	Penginapan umum
Anwar (2022)	*Instagram*	81%	Instagram, bukan *TikTok*

3. METODE PENELITIAN

3.1. Kerangka Penelitian

Penelitian ini menggunakan kerangka *kerja CRISP-DM (Cross-Industry Standard Process for Data Mining)* yang terdiri atas enam tahap: *Business Understanding*, Data *Understanding*, Data *Preparation, Modeling, Evaluation*, dan *Deployment*.



Gambar 1. Flowchart Alur Penelitian

3.2. Pengumpulan Data

Data dikumpulkan dari komentar *TikTok* terkait empat hotel di Kota Singkawang menggunakan layanan *ExportComments.com*, dihimpun dari 32 file CSV mentah sebelum melalui tahap *preprocessing*, dengan rincian pada Tabel 2.

Tabel 2. Spesifikasi Dataset

Hotel	File	Komentar
Hotel Mahkota Singkawang	8	451
Swiss-Belhotel Singkawang	8	269
Dayang Resort Singkawang	8	883
Horison Ultima Singkawang	8	663
Total	32	2.266

3.3. *Text Preprocessing*

Data mentah diproses melalui tujuh tahap berurutan: (1) *Emoji Handling* dan *Cleaning*, (2) *Case Folding*, (3) Normalisasi Karakter Berulang, (4) Normalisasi Kata, (5) *Stemming* menggunakan *library PySastrawi* dengan algoritma *Enhanced Confix Stripping* (ECS), (6) *Stopword Removal* dengan pengecualian kata negasi, dan (7) *Tokenizing*. Hasil rekapitulasi *preprocessing* ditunjukkan pada Tabel 3.

Tabel 3. Rekapitulasi Data *Preparation* per Hotel

Hotel	Asal	Final	Reduksi
Mahkota Singkawang	451	349	22,6%
Swiss-Belhotel	269	269	0%
Dayang Resort	883	699	20,8%
Horison Ultima	663	441	33,5%
Total	2.266	1.758	22,4%

3.4. Pelabelan Data dengan *Hybrid Lexicon*

Pelabelan sentimen menerapkan pendekatan *Hybrid Lexicon*: setiap token dicocokkan dengan *InSet Lexicon* yang telah diperluas dengan leksikon domain hotel, dengan kata ambigu dan kata topikal netral difilter agar tidak menyumbang skor yang menyesatkan. Penanganan negasi berbasis *window* (maksimal dua token) diterapkan agar frasa seperti "tidak nyaman" dapat terdeteksi sebagai sentimen negatif. Skor akhir kalimat dijumlahkan dari seluruh token, dan diberlakukan ambang bukti minimum $|\text{skor}| \geq 2$ agar kalimat berbukti lemah dikategorikan Netral.

3.5. Ekstraksi Fitur dan Pemodelan

Data berlabel dibagi menjadi 80% data latih dan 20% data uji secara *stratified*. Teks diekstraksi menjadi representasi numerik menggunakan *TF-IDF* dengan proses fitting hanya pada data latih untuk mencegah data *leakage*, kemudian model *Multinomial Naïve Bayes* dilatih dengan parameter alpha = 1,0 sebagai Laplace Smoothing.

4. HASIL DAN PEMBAHASAN

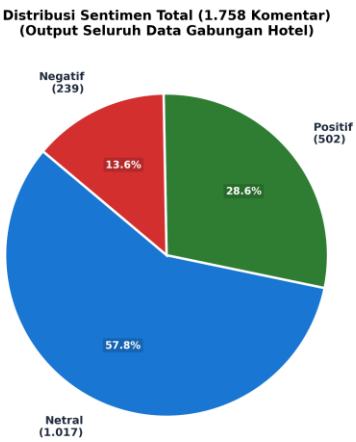
4.1. Distribusi Sentimen

Hasil pelabelan *Hybrid Lexicon* terhadap 1.758 komentar bersih menunjukkan kelas Netral mendominasi sebesar 57,8% (1.017 komentar),

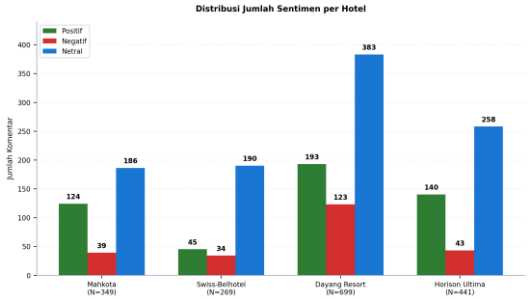
diikuti Positif 28,6% (502 komentar) dan Negatif 13,6% (239 komentar), sebagaimana dirinci pada Tabel 4 dan divisualisasikan pada Gambar 2, Gambar 3, dan Gambar 4. Dominasi kelas Netral merupakan konsekuensi langsung penerapan ambang bukti minimum, yang secara sengaja mengategorikan kalimat berbukti sentimen lemah sebagai Netral guna meningkatkan presisi pelabelan.

Tabel 4. Distribusi Sentimen *Hybrid Lexicon* per Hotel

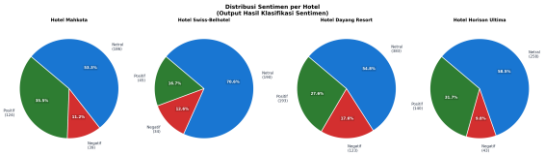
Hotel	Positif	Negatif	Netral	Total
Mahkota	124 (35,5%)	39 (11,2%)	186 (53,3%)	349
Swiss-Belhotel	45 (16,7%)	34 (12,6%)	190 (70,6%)	269
Dayang Resort	193 (27,6%)	123 (17,6%)	383 (54,8%)	699
Horison Ultima	140 (31,7%)	43 (9,8%)	258 (58,5%)	441
Total	502 (28,6%)	239 (13,6%)	1.017 (57,8%)	1.758



Gambar 2. Proporsi Sentimen Keseluruhan



Gambar 3. Distribusi Sentimen per Hotel



Gambar 4. Proporsi Sentimen per Hotel

4.2. Pembagian Data dan Ekstraksi Fitur

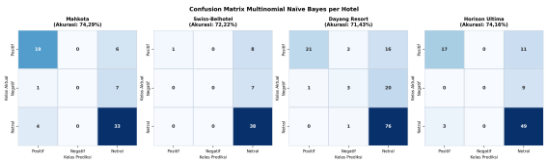
Dari 1.758 komentar berlabel, data dibagi menjadi 1.405 data latih dan 353 data uji secara *stratified*, dengan rincian per hotel pada Tabel 5.

Tabel 5. Pembagian Data Latih dan Uji

Hotel	Total	Latih	Uji
Mahkota	349	279	70
Swiss-Belhotel	269	215	54
Dayang Resort	699	559	140
Horison Ultima	441	352	89
Total	1.758	1.405	353

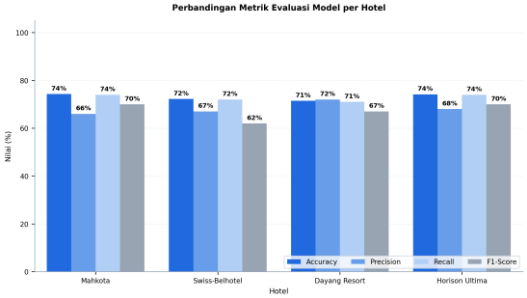
4.3. Hasil Evaluasi Model *Multinomial Naïve Bayes*

Model dievaluasi menggunakan confusion matrix serta metrik accuracy, *precision*, *recall*, dan *F1-score* pada tiap hotel, sebagaimana divisualisasikan pada Gambar 5.

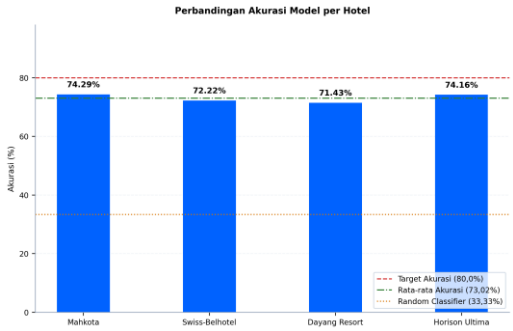


Gambar 5. *Confusion Matrix* Model *Multinomial Naïve Bayes* per Hotel

Hotel Mahkota Singkawang memperoleh akurasi tertinggi bersama Horison Ultima, yaitu 74,29%, sedangkan Dayang Resort memperoleh akurasi terendah sebesar 71,43%. Pola pada *confusion matrix* menunjukkan model relatif baik mengenali kelas Netral yang dominan, namun masih mengalami kesulitan pada kelas Negatif dengan jumlah data lebih sedikit, khususnya pada Hotel Swiss-Belhotel.



Gambar 6. Perbandingan Metrik Evaluasi Antar Hotel



Gambar 7. Perbandingan Akurasi Antar Hotel

4.4. Rekapitulasi Evaluasi dan Analisis Ketidakseimbangan Kelas

Rekapitulasi metrik evaluasi seluruh hotel disajikan pada Tabel 6. Rata-rata accuracy yang diperoleh sebesar 73,02%, *precision* 0,68, *recall* 0,73, dan *F1-score* 0,67.

Tabel 6. Rekapitulasi Metrik Evaluasi (Weighted Average)

Hotel	Akurasi	*Precision*	*Recall*	*F1-score*
Mahkota	74,29%	0,66	0,74	0,70
Swiss-Belhotel	72,22%	0,67	0,72	0,62
Dayang Resort	71,43%	0,72	0,71	0,67
Horison Ultima	74,16%	0,68	0,74	0,70
Rata-rata	73,02%	0,68	0,73	0,67

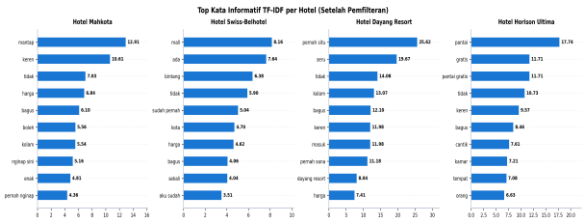
Rasio kelas Positif terhadap Negatif pada data keseluruhan adalah 2,10:1, jauh lebih seimbang dibandingkan pendekatan pelabelan leksikon murni pada iterasi penelitian sebelumnya yang mencapai rasio di atas 3:1. Rata-rata *accuracy* 73,02% berada 39,7 poin persentase di atas baseline random classifier tiga kelas (33,3%), menunjukkan model berhasil mempelajari pola sentimen secara bermakna.

4.5. Analisis Fitur *TF-IDF*

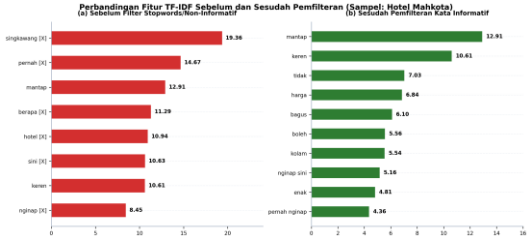
Analisis bobot *TF-IDF* pada data latih setelah tahap *filtering* menunjukkan fitur khas yang berbeda pada tiap hotel, sebagaimana dirangkum pada Tabel 7 dan divisualisasikan pada Gambar 8 dan Gambar 9.

Tabel 7. Top-5 Fitur *TF-IDF* per Hotel Setelah Filter

No	Mahkota	Swiss-Belhotel	Dayang Resort	Horison Ultima
1	mantap (12,91)	mall (8,16)	pernah situ (25,62)	pantai (17,74)
2	keren (10,61)	ada (7,64)	seru (19,67)	gratis (11,71)
3	tidak (7,03)	bintang (6,38)	tidak (14,08)	pantai gratis (11,71)
4	harga (6,84)	tidak (5,90)	kolam (13,07)	tidak (10,73)
5	bagus (6,10)	harga (4,62)	bagus (12,10)	keren (9,57)



Gambar 8. Heatmap *TF-IDF* Setelah Filter



Gambar 9. Perbandingan Fitur Sebelum dan Sesudah Filter *Hybrid Lexicon*

Kata "mantap" dan "keren" menjadi fitur dominan pada Hotel Mahkota, "mall" pada Swiss-Belhotel mengindikasikan lokasi strategis, "seru" dan "kolam" pada Dayang Resort mencerminkan daya tarik wahana air, sementara "pantai" dan "gratis" pada Horison Ultima menegaskan lokasi tepi pantai sebagai nilai jual utama.

4.6. Keterbatasan Penelitian

Penelitian ini memiliki beberapa keterbatasan: (1) jumlah data uji pada Hotel Swiss-Belhotel (54 data) relatif kecil sehingga evaluasi pada kelas minoritas kurang stabil, (2) leksikon domain hotel serta kosakata slang lokal Kalimantan Barat masih terbatas cakupannya, dan (3) dominasi kelas Netral akibat ambang bukti minimum berpotensi mengurangi sensitivitas model terhadap sentimen yang terekspresikan secara implisit.

4.7. Pembahasan

Dibandingkan dengan pendekatan pelabelan leksikon murni pada iterasi penelitian sebelumnya yang menghasilkan rata-rata *accuracy* berkisar 61–64% dengan dominasi kelas Positif, pendekatan *Hybrid Lexicon* pada penelitian ini menghasilkan rata-rata accuracy yang lebih tinggi (73,02%) dengan distribusi kelas yang lebih proporsional. Hal ini menunjukkan bahwa penambahan filter kata ambigu, filter topikal netral, leksikon domain, serta ambang bukti minimum secara efektif mengurangi kesalahan pelabelan yang sebelumnya menyebabkan bias berlebihan terhadap kelas Positif.

5. KESIMPULAN

Berdasarkan hasil penelitian, dapat disimpulkan dua hal utama. Pertama, dari 2.266 komentar mentah yang dihimpun melalui 32 file CSV, diperoleh 1.758 komentar bersih (reduksi 22,4%) setelah tujuh tahap *preprocessing*. Pelabelan dengan pendekatan *Hybrid Lexicon* menghasilkan distribusi Netral 57,8%, Positif 28,6%, dan Negatif 13,6%, dengan rasio Positif:Negatif 2,10:1. Kedua, model *Multinomial Naïve Bayes* dengan pembobotan *TF-IDF* menghasilkan rata-rata accuracy 73,02%, *precision* 0,68, *recall* 0,73, dan *F1-score* 0,67, unggul 39,7 poin persentase di atas baseline random classifier (33,3%), membuktikan pendekatan *Hybrid Lexicon* berkontribusi nyata terhadap peningkatan kualitas pelabelan dan performa klasifikasi.

Saran untuk penelitian selanjutnya meliputi: (1) perluasan leksikon domain hotel dan kosakata slang lokal Kalimantan Barat, (2) penambahan jumlah data uji khususnya pada hotel dengan data terbatas, (3) perbandingan dengan algoritma lain seperti SVM atau IndoBERT, dan (4) evaluasi lanjutan terhadap ambang bukti minimum.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada dosen pembimbing dan dosen penguji atas bimbingan dan masukan yang diberikan selama proses penelitian dan penyusunan Tugas Akhir ini, serta kepada Politeknik Negeri Sambas atas dukungan fasilitas yang diberikan.

DAFTAR PUSTAKA

AZLAN, R. & RAHMI, F., 2025. Analisis sentimen ulasan penginapan pada Google Review menggunakan Naïve Bayes. Jurnal Ilmiah Informatika, 10(1).

ANWAR, S., 2022. Analisis sentimen komentar Instagram menggunakan Naïve Bayes Classifier. Jurnal Computer Science, 3(2).

AZZAHRA, N. dkk., 2024. Analisis sentimen ulasan aplikasi *TikTok* pada Google Play Store. Jurnal Ilmu Komputer, 8(1).

HIKMAL, M.F., 2023. Perkembangan pariwisata dan media sosial sebagai sumber opini wisatawan. Jurnal Pariwisata Digital, 5(2).

KOTO, F. & MAHENDRA, G.Y., 2017. Inset lexicon: Evaluation of a word list for Indonesian sentiment analysis in microblogs. Dalam: Proc. 2017 Int. Conf. Asian Lang. Process. (IALP).

LIU, B., 2020. Sentiment Analysis: Mining Opinions, Sentiments, and Emotions. 2nd ed. Cambridge: Cambridge University Press.

MCCALLUM, A. & NIGAM, K., 1998. A comparison of event models for Naive Bayes text classification. Dalam: AAAI-98 Workshop on Learning for Text Categorization, pp.41–48.

MEDHAT, W., HASSAN, A. & KORASHY, H., 2014. Sentiment analysis algorithms and applications: A survey. Ain Shams Engineering Journal, 5(4), pp.1093–1113.

RIFA'I, M. dkk., 2021. Analisis sentimen ulasan wisata Kalimantan Barat pada Google Maps. Jurnal Teknologi Informasi, 4(1).

SALTON, G. & BUCKLEY, C., 1988. Term-weighting approaches in automatic text retrieval. Information Processing & Management, 24(5), pp.513–523.

THOMAS, A. & RUMAISA, R., 2022. Klasifikasi sentimen ulasan hotel pada Tripadvisor menggunakan Support Vector Machine. Jurnal Sistem Informasi, 6(2).