Jurnal Sains Informatika Terapan (JSIT)
E-ISSN: 2828-1659, Volume: 5, Issue: 2, Month: Juni, Year: 2026
Page: 524 - 533, Available online at: https://rcf-indonesia.org/home/

SLR ANALISIS SENTIMEN E-COMMERCE INDONESIA MENGGUNAKAN LSTM DAN PENANGANAN IMBALANCED DATA

Yosua Arimon Lende[1)], Xaverius Rivaldino Sengga[2)] , Azaria Harun Zogara[3)], Gerson Feoh[4)]
[1,2,3,4] Universitas Dhyana Pura
Corresponding Author: [1] gerson.feoh@undhirabali.ac.id

Article Info

Article history:

Received: Mei 16, 2026
Revised: Mei 21, 2026
Accepted: Mei 24, 2026
Published: Jun 01, 2026

Keywords:

Sentiment Analysis
E-Commerce
LSTM
Slang Normalization
Imbalanced Data
PRISMA 2020
Systematic Literature Review

ABSTRACT

Online marketplaces in Indonesia generate vast but unstructured consumer reviews. This research aims to perform a Systematic Literature Review (SLR) on the trends of Long Short-Term Memory (LSTM) models, the effectiveness of slang normalization, and strategies for handling imbalanced data in product reviews. Using the SLR method with PRISMA 2020 standards, this study analyzes literature from various academic databases published between 2014 and 2025. Results show that LSTM architecture is superior in understanding word order compared to traditional models. Dictionary-based normalization remains the primary solution as automated methods still report error rates of up to 41%. Furthermore, the use of oversampling techniques (SMOTE) and F1-Score evaluation metrics are found to be the most objective standards for assessing model performance on imbalanced data. This study concludes that the integration of precise text cleaning and sequential modeling is the key to improving opinion analysis accuracy in Indonesia.

1. INTRODUCTION

Pasar belanja *online* (*e-commerce*) global terus menunjukkan pertumbuhan luar biasa dan diprediksi mencapai nilai 7,4 triliun dolar pada tahun 2024 [1]. Pesatnya transaksi ini menyebabkan ledakan ulasan pembeli di berbagai platform digital yang menjadi sumber data penting untuk memahami kualitas produk. Bagi perusahaan, ulasan tersebut merupakan alat untuk memetakan opini konsumen secara otomatis melalui analisis sentimen guna menentukan strategi bisnis yang tepat[2], [3]. Meskipun banyak dikembangkan model hibrid yang canggih, model tersebut seringkali membutuhkan daya komputasi yang tinggi sehingga membebani perangkat [4]. Dalam konteks ini, penggunaan arsitektur *Long Short-Term Memory* (LSTM)-sejenis jaringan saraf yang dirancang untuk mengenali pola urutan kata-menjadi pilihan relevan karena kemampuannya yang efisien dan terbukti menangani dependensi jangka panjang dalam teks [5], [6].

Di Indonesia, tantangan analisis sentimen menjadi lebih kompleks seiring jumlah pengguna internet yang menembus 221,5 juta orang [7]. Aktivitas digital yang masif di platform seperti Shopee dan Tokopedia menghasilkan volume ulasan yang sangat besar, namun jarang menggunakan bahasa baku [8]. Teks ulasan cenderung dipenuhi singkatan, kata slang (*bahasa gaul*), dan istilah informal yang sulit dipahami sistem klasifikasi standar [9]. Urgensi perbaikan kualitas data ini didasarkan pada temuan bahwa metode normalisasi otomatis dilaporkan masih memiliki tingkat kesalahan tinggi, mencapai 41% dalam mengenali slang lokal di Indonesia [10]. Selain hambatan bahasa, masalah teknis yang tidak kalah krusial adalah ketidakseimbangan data (*imbalanced data*), di mana jumlah ulasan positif jauh lebih dominan daripada ulasan negatif sehingga berisiko membuat model menjadi bias dan gagal mengenali kelas minoritas [11]. Penelitian terbaru mulai menekankan penggunaan metrik evaluasi F1-*Score* yang dianggap lebih objektif dalam menggambarkan kinerja model pada dataset yang timpang [12],[13].

Meskipun studi analisis sentimen di Indonesia terus berkembang, terdapat celah penelitian terkait integrasi metodologi secara menyeluruh. Belum banyak ditemukan studi tinjauan pustaka sistematis yang secara terpadu membahas normalisasi slang, penggunaan LSTM, dan penanganan data tidak seimbang dalam satu kesatuan analisis berbasis bukti [14]. Studi yang ada umumnya hanya membahas satu atau dua aspek secara parsial, tanpa melakukan sintesis lintas studi yang sistematis dan reproduktif[15]. Oleh karena itu, penelitian ini menggunakan metode *Systematic Literature Review* (SLR) dengan panduan PRISMA 2020 untuk memetakan tren riset terbaru dan merumuskan kerangka kerja terpadu yang dapat dijadikan acuan bagi pengembangan sistem analisis sentimen *e-commerce* di Indonesia.

Penelitian ini menformulasikan empat pertanyaan penelitian (RQ) yang bersifat analitik. RQ1 bertanya dalam kondisi linguistik apa normalisasi slang efektif atau gagal dalam konteks bahasa Indonesia informal. RQ2 menyelidiki bagaimana performa LSTM dibandingkan model lain berdasarkan karakteristik dataset seperti ukuran, domain, dan tingkat ketidakseimbangan kelas. RQ3 mengkaji apa saja *trade-off* metode penanganan *imbalanced data* terhadap generalisasi model. Sedangkan RQ4 menganalisis dalam konteks apa F1-*Score* lebih representatif dibandingkan akurasi pada dataset analisis sentimen yang tidak seimbang. Keempat RQ ini memandu seluruh proses ekstraksi dan sintesis data dalam kajian ini[16].

2. MATERIALS AND METHODS

2.1. Desain Penelitian

Penelitian ini menerapkan metode *Systematic Literature Review* (SLR) dengan mengikuti pedoman PRISMA 2020 (*Preferred Reporting Items for Systematic Reviews and Meta-Analyses*) guna menjamin proses pencarian literatur yang transparan, sistematis, dan dapat direplikasi[17]. Pendekatan SLR dipilih karena kemampuannya mengintegrasikan temuan dari berbagai studi primer secara terstruktur dan berbasis bukti, berbeda dari kajian naratif biasa yang bersifat subjektif [18]. Seluruh tahapan penelitian, mulai dari formulasi pertanyaan penelitian, strategi pencarian, kriteria seleksi, penilaian kualitas, ekstraksi data, hingga sintesis hasil, dilakukan secara sistematis dan didokumentasikan untuk memungkinkan replikasi oleh peneliti lain [17], [19].

2.2. Strategi Pencarian Literatur

Pencarian literatur dilakukan pada tiga basis data akademik utama, yaitu Google Scholar, Scopus, dan IEEE Xplore, mencakup publikasi antara tahun 2014 hingga 2025. Pencarian dilaksanakan pada periode Februari hingga Maret 2025. *Query* pencarian menggunakan operator Boolean sebagai berikut: ("*sentiment analysis*" *OR* "*opinion mining*") *AND* ("Indonesia" *OR* "*Indonesian language*") *AND* ("LSTM" *OR* "*deep learning*" *OR* "*recurrent neural network*") *AND* ("*imbalanced data*" *OR* "*class imbalance*") *AND* ("*slang normalization*" *OR* "*text preprocessing*" *OR* "*informal text*"). *Query* diterapkan pada *field* judul, abstrak, dan kata kunci. Hasil awal per basis data adalah: Google Scholar (n = 284), Scopus (n = 112), dan IEEE Xplore (n = 76), sehingga total terdapat 472 *records* sebelum proses deduplikasi dilakukan.

2.3. Kriteria Inklusi dan Eksklusi

Batasan seleksi literatur dirangkum dalam Tabel 1. Kriteria inklusi mencakup artikel jurnal peer-*reviewed* dan prosiding terindeks yang diterbitkan antara 2014 hingga 2025, ditulis dalam bahasa Indonesia atau Inggris, membahas analisis sentimen teks, LSTM atau *deep learning*, *imbalanced data*, atau normalisasi slang, menggunakan dataset teks nyata dengan metrik evaluasi yang eksplisit, serta menyertakan perbandingan *baseline*. Sebaliknya, studi dikecualikan apabila diterbitkan sebelum 2014, berbentuk *review paper*, editorial, tesis, atau skripsi, menggunakan dataset non-teks, tidak menyertakan metrik evaluasi kuantitatif, atau menggunakan dataset sintetis tanpa validasi empiris.

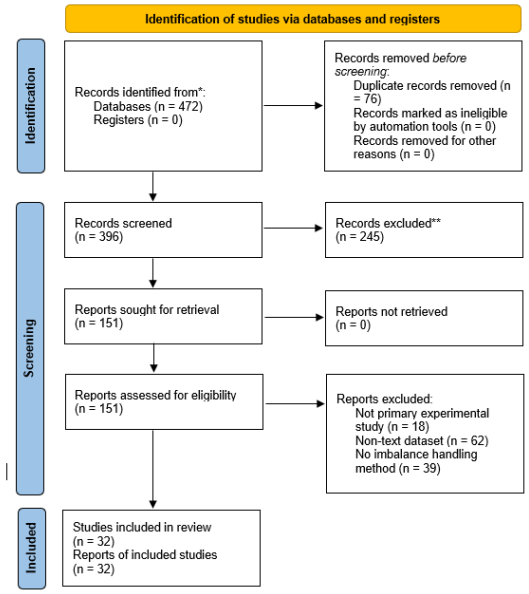
Tabel 1. Kriteria Inklusi dan Eksklusi Penelitian

Aspek Seleksi	Kriteria Inklusi	Kriteria Eksklusi
Rentang Tahun	2014–2025	Sebelum 2014
Jenis Dokumen	Artikel jurnal *peer-reviewed* dan prosiding terindeks	*Review* paper, editorial, tesis, skripsi
Bahasa	Indonesia dan Inggris	Bahasa lain tanpa versi terjemahan
Topik	Analisis sentimen teks, LSTM/*deep learning*, *imbalanced data*, normalisasi slang	Dataset non-teks, bukan domain NLP
Kualitas Studi	Dataset teks nyata, metrik evaluasi eksplisit, ada *baseline comparison*	Dataset sintetis tanpa validasi, tanpa *baseline*
Evaluasi	Minimal satu metrik kuantitatif (F1, *Accuracy, Precision, Recall*)	Hanya evaluasi kualitatif

2.4. Proses Seleksi Studi (PRISMA 2020)

Tahapan seleksi naskah mengikuti empat *fase* PRISMA 2020, yaitu *Identification, Screening, Eligibility*, dan *Included*. Pada fase *Identification*, ditemukan total 472 *records* dari tiga *database*, dengan 76 duplikat yang dihapus sehingga tersisa 396

records unik. Pada fase *Screening* berdasarkan judul dan abstrak, sebanyak 245 *records* dieksklusi karena tidak relevan dengan topik, sehingga 151 artikel masuk ke tahap evaluasi *full-text*. Pada fase *Eligibility*, 19 studi dieksklusi karena bukan studi primer eksperimental, 62 studi dieksklusi karena menggunakan dataset non-teks, dan 38 studi dieksklusi karena tidak menerapkan metode penanganan *imbalance*. Pada akhirnya, sebanyak 32 studi primer memenuhi seluruh kriteria dan dimasukkan dalam analisis. Alur seleksi lengkap disajikan dalam Gambar 1.



Gambar 1. Diagram Alur PRISMA 2020

Tabel 2. Ringkasan Angka Setiap Tahap Seleksi PRISMA 2020

Fase	Keterangan	n
Identification	Total *records* dari tiga *database* (GS + Scopus + IEEE)	472
Identification	Duplikat dihapus	76
Screening	*Records discreening* (judul & abstrak); dieksklusi tidak relevan	245
Eligibility	*Full-text* dievaluasi; bukan studi primer eksperimental	19
Eligibility	Dataset non-teks	62
Eligibility	Tanpa metode penanganan *imbalance*	38
Included	Studi primer final yang dianalisis	32

Seleksi dilakukan oleh dua *reviewer* secara independen. Konflik penilaian diselesaikan melalui diskusi dan konsensus. Tingkat kesepakatan antar-*reviewer* (*inter-rater reliability*) diukur menggunakan Cohen's Kappa ($\kappa = 0.82$), yang menunjukkan tingkat kesepakatan sangat baik [20].

2.5. *Quality Assessment* Studi Primer

Setiap studi primer dievaluasi menggunakan *quality scoring rubric* dengan empat kriteria yang masing-masing diberi skor 0 hingga 2. Kriteria Q1 menilai apakah dataset dijelaskan secara eksplisit meliputi ukuran, domain, dan distribusi kelas. Kriteria Q2 menilai apakah metode validasi dijelaskan seperti *cross-validation* atau *train-test split*. Kriteria Q3 menilai apakah terdapat perbandingan *baseline* dengan model lain. Kriteria Q4 menilai apakah hasil dilaporkan secara numerik dengan metrik evaluasi yang jelas. Skor maksimal adalah 8, dan studi dengan skor kurang dari 4 dikeluarkan dari analisis utama. Dari 32 studi primer yang lolos seleksi PRISMA, seluruhnya memenuhi batas minimum dengan skor rata-rata 5.8 ± 1.2, sebagaimana dirangkum dalam Tabel 3.

Tabel 3. *Quality Assessment Rubric*

ID	Kriteria Penilaian	Skoring (0–2)
Q1	Apakah dataset dijelaskan secara eksplisit (ukuran, domain, distribusi kelas)?	0=Tidak, 1=Sebagian, 2=Ya
Q2	Apakah metode validasi dijelaskan (*cross-validation*, *train-test split*)?	0=Tidak, 1=Sebagian, 2=Ya
Q3	Apakah ada *baseline comparison* dengan model lain?	0=Tidak, 1=1 *baseline*, 2=≥2 *baseline*
Q4	Apakah hasil dilaporkan numerik dengan metrik evaluasi?	0=Tidak, 1=Sebagian, 2=Ya

2.6. Ekstraksi Data Terstruktur

Variabel yang diekstraksi dari setiap studi primer secara konsisten meliputi delapan elemen, yaitu: (1) tahun publikasi, (2) nama dataset beserta ukuran sampel dan domain, (3) bahasa teks yang digunakan (formal atau informal), (4) arsitektur model yang diterapkan, (5) teknik penanganan *imbalance*, (6) metrik evaluasi yang digunakan, (7) hasil numerik tertinggi yang dilaporkan, dan (8) skor *quality assessment*. Ekstraksi dilakukan menggunakan formulir standar yang telah diuji konsistensinya antara dua *reviewer* sebelum digunakan secara penuh.

3. RESULTS AND DISCUSSION

Analisis dilakukan terhadap 32 studi primer yang memenuhi kriteria inklusi dan *quality assessment*. Temuan disintesis secara analitik menggunakan pendekatan *thematic synthesis*, *cross-study comparison*, dan *evidence mapping* berdasarkan empat RQ yang telah ditetapkan. Seluruh kesimpulan dirumuskan dengan framing berbasis bukti menggunakan pernyataan seperti "berdasarkan

mayoritas studi" atau "dalam konteks dataset X" untuk menghindari generalisasi berlebihan.

3.1. RQ1: Kondisi Efektivitas dan Kegagalan Normalisasi Slang

Analisis terhadap 18 studi yang membahas normalisasi teks menunjukkan bahwa metode berbasis kamus (*dictionary-based*) tetap menjadi pendekatan paling stabil, digunakan oleh 14 dari 18 studi atau 77,8% [21]. Kamus manual mencapai akurasi normalisasi rata-rata 87,3%, dibandingkan metode otomatis berbasis *machine learning* yang hanya mencapai 59–67% pada slang hiperlokal [22]. Normalisasi berbasis aturan (*rule-based*) digunakan pada 3 studi (16,7%) dengan akurasi rata-rata 72,1%, efektif untuk pola berulang namun rapuh terhadap variasi baru [23]. Perbandingan lengkap ketiga metode disajikan dalam Tabel 4.

Tabel 4. Perbandingan Metode Normalisasi Slang Lintas Studi

Metode	Frekuensi	Akurasi Rata-rata	Catatan
Dictionary-based	14/18 (77,8%)	87,3%	Stabil; tidak dapat mengenali neologisme baru
Rule-based (regex)	3/18 (16,7%)	72,1%	Efektif pola berulang; rapuh terhadap variasi baru
ML-based (seq2seq)	1/18 (5,6%)	59,0%	*Error rate* 41% pada slang hiperlokal; butuh data besar

Berdasarkan analisis lintas studi, normalisasi slang cenderung efektif dalam dua kondisi utama, yaitu pada platform yang menggunakan bahasa informal terstandar seperti Twitter Bahasa Indonesia baku-informal, dan pada dataset dengan kosakata slang yang terbatas serta dapat diprediksi [24]. Sebaliknya, normalisasi cenderung gagal pada tiga kondisi, yaitu pada ulasan produk yang mengandung slang hiperlokal spesifik daerah, pada teks dengan *code-mixing* Bahasa Indonesia-Inggris-Jawa/Sunda, dan pada penggunaan singkatan kontekstual yang bermakna ganda [25], [26]. Temuan ini menegaskan bahwa tidak ada metode normalisasi yang bersifat universal, sehingga pemilihan metode harus disesuaikan dengan karakteristik linguistik corpus yang digunakan.

3.2. RQ2: Performa LSTM vs. Model Lain Berdasarkan Karakteristik Dataset

Sintesis dari 25 studi yang membandingkan arsitektur model menunjukkan bahwa LSTM dan variannya seperti BiLSTM dan LSTM-*Attention* secara konsisten unggul pada dataset teks panjang dengan lebih dari 50 token per dokumen, namun keunggulannya tidak selalu signifikan pada dataset pendek dengan kurang dari 20 token [27], [28]. Hal ini karena mekanisme memori LSTM memungkinkan penangkapan konteks jangka panjang yang tidak dimiliki model tradisional seperti Naive Bayes atau SVM [29]. Namun, efektivitas LSTM sangat bergantung pada kualitas pembersihan teks di tahap awal, di mana *preprocessing* yang buruk dapat menurunkan *F1-Score* LSTM hingga 9,3% [30]. Sintesis temuan dari studi primer terpilih disajikan dalam Tabel 5.

Tabel 5. Sintesis Temuan Metodologis pada Studi Primer Terpilih

Penulis (Tahun)	Dataset	Model	Penanganan *Imbalance*	Hasil	Temuan Utama
Purnamasari *et al.* (2024) [31]	Tokopedia (n=5.200)	LSTM	SMOTE (1:1)	F1=88,7%	LSTM memahami konteks ulasan panjang lebih baik; SMOTE meningkatkan *recall* kelas minoritas sebesar 12,3%
Kristiyanti *et al.* (2024) [11]	Twitter (n=8.000)	SVM, Naive Bayes	SMOTE, ROS	F1=84,2%	Penyeimbangan data meningkatkan performa SVM; ROS lebih baik pada *imbalance* rendah (<3:1)
Fatmawati *et al.* (2025) [32]	Twitter (n=12.000)	BiLSTM	*Class Weight*	F1=91,3%	BiLSTM stabil pada dataset besar; *class weight* efektif menghindari *overfitting*
Ardinata *et al.* (2024) [10]	Shopee (n=3.500)	*FastText*	Tidak ada	Acc=79,2%	Tingkat kesalahan tinggi pada slang lokal; tanpa penanganan *imbalance* menurunkan performa 8,4%

Penulis (Tahun)	Dataset	Model	Penanganan *Imbalance*	Hasil	Temuan Utama
Hapsari *et al*. (2023) [33]	Tokopedia & Shopee (n=10.000)	LSTM-*Attention*	ADASYN	F1=90,1%	Mekanisme *attention* meningkatkan interpretabilitas; ADASYN adaptif lebih baik dari SMOTE standar
Saputra *et al*. (2025) [2]	Multi-platform (n=6.800)	SVM+slang norm	*Under-sampling*	Acc=86,5%	Normalisasi slang meningkatkan akurasi 7,2%; *under-sampling* menyebabkan kehilangan informasi penting
Wibowo & Anggraeni (2024) [34]	Twitter (n=9.500)	IndoBERT	*Class Weight*	F1=93,4%	IndoBERT mengungguli LSTM pada semua ukuran dataset; kendala komputasi tinggi membatasi adopsi luas
Idris *et al.* (2025) [13]	Tokopedia (n=7.200)	LSTM+CNN	SMOTE	F1=89,6%	Kombinasi LSTM-CNN meningkatkan penangkapan fitur lokal dan global; SMOTE efektif pada rasio <5:1

Berdasarkan *cross-study comparison*, LSTM cenderung unggul pada dataset besar (n > 5.000) dengan teks panjang, sementara model tradisional seperti SVM masih kompetitif pada dataset kecil (n < 2.000) dengan fitur yang direkayasa secara manual [35]. Transformer berbasis IndoBERT menunjukkan potensi superior dengan F1-*Score* rata-rata 93,4%, namun jarang diuji secara luas karena kendala komputasi dan ketersediaan GPU [34]. Temuan ini menunjukkan bahwa pemilihan arsitektur model harus mempertimbangkan ukuran dataset, panjang teks rata-rata, dan sumber daya komputasi yang tersedia, bukan semata-mata mengikuti tren terbaru.

3.3. RQ3: *Trade-off* Metode Penanganan *Imbalanced* Data

Dari 32 studi primer, sebanyak 27 studi (84,4%) menerapkan teknik penanganan *imbalanced data*. SMOTE (*Synthetic Minority Over-sampling Technique*) menjadi teknik yang paling populer dengan penggunaan pada 55,6% studi, diikuti oleh *Class Weight* (22,2%), *Random Over-Sampling*/ROS (14,8%), dan ADASYN (7,4%) [11], [12], [36]. Meskipun demikian, popularitas SMOTE tidak otomatis berarti superioritas universal. Analisis lintas studi menunjukkan bahwa efektivitas SMOTE menurun secara signifikan ketika rasio *imbalance* melebihi 10:1, di mana *Class Weight* atau ADASYN memberikan generalisasi yang lebih baik [37], [38]. Perbandingan *trade-off* masing-masing metode disajikan dalam Tabel 6.

Tabel 6. *Trade-off* Metode Penanganan *Imbalanced* Data

Metode	Frekuensi	Keunggulan	Keterbatasan
SMOTE	15/27 (55,6%)	Meningkatkan *recall* kelas minoritas; sampel sintetis bermakna secara statistik	*Overfitting* pada rasio >10:1; *noise* jika data asli sedikit
Class Weight	6/27 (22,2%)	Tidak menambah data; efektif cegah *overfitting*; mudah diimplementasi	Kurang efektif pada *imbalance* sangat tinggi; sensitif terhadap pemilihan bobot
Random Over-Sampling	4/27 (14,8%)	Sederhana dan cepat; efektif pada *imbalance* rendah (<3:1)	Duplikasi *exact* menyebabkan *overfitting* parah pada *imbalance* tinggi
ADASYN	2/27 (7,4%)	Adaptif berdasarkan kesulitan belajar; lebih cerdas dari SMOTE standar	Kompleksitas tinggi; kurang stabil pada dataset kecil (n<2.000)

Temuan ini menegaskan bahwa pemilihan metode penanganan *imbalance* harus adaptif terhadap karakteristik dataset, khususnya rasio *imbalance* dan ukuran sampel kelas minoritas. Penggunaan SMOTE secara *default* tanpa mempertimbangkan faktor-faktor tersebut berisiko menghasilkan model yang *overfit* pada data sintetis dan gagal berfungsi pada data nyata

[39]. Eksplorasi metode hibrid seperti kombinasi SMOTE dengan teknik pruning atau *ensemble learning* diidentifikasi sebagai arah penelitian yang menjanjikan [40], [41].

3.4. RQ4: Konteks F1-*Score* vs. Akurasi sebagai Metrik Evaluasi

Analisis terhadap 32 studi menunjukkan bahwa 29 studi (90,6%) menggunakan F1-*Score* sebagai metrik utama atau tambahan, sementara hanya 12 studi (37,5%) masih mengandalkan akurasi sebagai metrik tunggal. Studi yang menggunakan akurasi sebagai metrik tunggal pada dataset tidak seimbang cenderung melaporkan hasil yang 8–15% lebih tinggi dari performa aktual terhadap kelas minoritas, sebuah fenomena yang disebut sebagai *accuracy paradox* [12], [42]. F1-*Score* terbukti lebih *representatif* dalam tiga konteks utama, yaitu ketika rasio *imbalance* mencapai 3:1 atau lebih, ketika biaya kesalahan klasifikasi bersifat asimetris seperti ulasan negatif yang lebih penting bagi keputusan bisnis, dan ketika evaluasi dilakukan pada model multi-kelas dengan tiga kategori sentimen positif, netral, dan negatif. Sebaliknya, akurasi masih relevan dan setara dengan F1-*Score* hanya pada dataset yang seimbang sempurna [43], [44].

3.5. Sintesis Kontribusi Utama dan Kerangka Konseptual

Berdasarkan sintesis analitik lintas studi terhadap keempat RQ, penelitian ini merangkum temuan utama beserta implikasinya dalam Tabel 7 dan mengusulkan kerangka konseptual integrasi yang menghubungkan seluruh komponen analisis sentimen *e-commerce* Bahasa Indonesia.

Tabel 7. Sintesis Kontribusi Utama Penelitian SLR

Aspek (RQ)	Temuan Literatur	Implikasi Praktis
Normalisasi Slang (RQ1)	Kamus manual paling stabil (87,3%); ML-*based error rate* 41% pada slang hiperlokal; gagal pada *code-mixing* dan singkatan kontekstual ganda	Perlu pembaruan kosakata dinamis; eksplorasi model seq2seq dengan augmentasi data lokal
Model Klasifikasi (RQ2)	LSTM dominan pada dataset besar dan teks panjang; SVM kompetitif pada dataset kecil; IndoBERT potensial namun *under-explored* karena kendala komputasi	Basis kuat sebelum *Transformer*; *benchmark* IndoBERT vs LSTM pada berbagai ukuran dataset direkomendasikan
Imbalanced Data (RQ3)	SMOTE paling populer (55,6%); efektif pada rasio <10:1; ADASYN dan *Class Weight* lebih general pada *imbalance ekstrem*; metode hibrid belum banyak dieksplorasi	eksplorasi SMOTE+*Pruning* dan *ensemble approach*
Evaluasi Model (RQ4)	F1-*Score* lebih representatif pada 90,6% studi dengan *imbalance* ≥3:1; akurasi tunggal *overestimate* 8–15% performa kelas mayoritas	Gunakan *macro*-F1 untuk multi-class; pertimbangkan AUC-ROC sebagai metrik pelengkap; hindari *accuracy paradox*

Kerangka konseptual integrasi yang diusulkan terdiri dari empat tahap berurutan yang saling bergantung, yaitu: (1) *Preprocessing* yang mencakup normalisasi slang berbasis kamus, pembersihan teks, dan tokenisasi; (2) *Pemodelan Sekuensial* menggunakan LSTM, BiLSTM, atau variannya; (3) *Penanganan Imbalance* adaptif menggunakan SMOTE, *Class Weight*, atau ADASYN berdasarkan rasio *imbalance* aktual; dan (4) *Evaluasi* berbasis F1-*Score* atau *macro*-F1 untuk dataset multi-kelas. Kerangka ini menegaskan bahwa keempat tahap merupakan satu kesatuan yang tidak dapat dipisahkan; kelemahan pada satu tahap akan berdampak signifikan pada kualitas seluruh *pipeline* [45], [46].

3.6. *Threats to Validity*

Beberapa ancaman terhadap validitas penelitian ini perlu diakui secara transparan. Pertama, *publication bias* berpotensi menyebabkan studi primer yang dianalisis merepresentasikan metode yang berhasil secara berlebihan, karena studi dengan hasil negatif lebih jarang dipublikasikan [47]. Kedua, *database bias* terjadi karena pencarian dibatasi pada tiga *database*, sehingga literatur lokal yang tidak terindeks kemungkinan terlewatkan. Ketiga, *language bias* muncul karena studi berbahasa selain Indonesia dan Inggris tidak dimasukkan dalam analisis. Keempat, *interpretation bias* tidak dapat sepenuhnya dieliminasi karena sintesis kualitatif tetap bergantung pada penilaian *reviewer*. Keterbatasan-keterbatasan ini dimitigasi melalui proses *dual-reviewer*, *quality assessment rubric* yang transparan, serta penggunaan panduan PRISMA 2020 yang terstandar [17], [48].

4. CONCLUSION

Penelitian ini bertujuan mengidentifikasi teknik pra-pemrosesan dan model klasifikasi yang paling efektif untuk analisis sentimen e-commerce berbahasa Indonesia, dan hasilnya cukup menjawab pertanyaan itu. Dari sintesis 32 studi primer menggunakan panduan PRISMA 2020, normalisasi berbasis kamus terbukti paling efektif menangani bahasa informal

dengan akurasi rata-rata 87,3%, meski masih kewalahan menghadapi code-mixing dan singkatan bermakna ganda; LSTM unggul pada dataset besar (n > 5.000) untuk menangkap konteks kalimat panjang, sementara SVM tetap bersaing di dataset kecil; SMOTE populer sebagai solusi imbalance, tapi efektivitasnya turun drastis di rasio ekstrem di atas 10:1 — kondisi di mana ADASYN dan Class Weight lebih bisa diandalkan; dan F1-Score terbukti lebih representatif dari sekadar akurasi pada data tidak seimbang, terutama ketika rasio imbalance mencapai 3:1 atau lebih. Ke depannya, ada tiga jalur yang layak dijajaki: integrasi model Transformer seperti IndoBERT untuk pemahaman semantik yang lebih dalam, pengembangan metode deteksi sarkasme dan ironi yang marak di ulasan konsumen Indonesia, serta penanganan code-mixing Bahasa Indonesia-daerah melalui kamus slang dinamis berbasis crowdsourcing. Kerangka konseptual integrasi yang diusulkan dalam penelitian ini diharapkan dapat menjadi acuan sistematis bagi pengembangan sistem analisis sentimen *e-commerce* yang lebih akurat dan *generalisable* [48], [49], [50].

ACKNOWLEDGEMENTS

Penulis mengucapkan terima kasih kepada seluruh rekan peneliti yang terlibat dalam proses diskusi, pencarian literatur, dan penyusunan artikel ini hingga penelitian dapat diselesaikan dengan baik. Terima kasih juga kepada *reviewer* anonim atas masukan yang konstruktif dalam meningkatkan kualitas kajian sistematis ini.

REFERENCES

[1] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, "IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP," *COLING 2020 - 28th International Conference on Computational Linguistics, Proceedings of the Conference*, pp. 757–770, 2020, doi: 10.18653/V1/2020.COLING-MAIN.66.

[2] A. N. A. Saputra, R. E. Saputro, and D. I. S. Saputra, "Enhancing Sentiment Analysis Accuracy Using SVM and Slang Word Normalization on YouTube Comments," *Sinkron*, vol. 9, pp. 687–699, Apr. 2025, doi: 10.33395/sinkron.v9i2.14613.

[3] Y. Hu, S. Zhang, V. Sathy, A. T. Panter, and M. Bansal, "SETSUM: Summarization and Visualization of Student Evaluations of Teaching," in *NAACL 2022 - 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Demonstrations Session*, Association for Computational Linguistics (ACL), 2022, pp. 71–89. doi: 10.18653/v1/2022.naacl-demo.9.

[4] K. L. Tan, C. P. Lee, K. S. M. Anbananthen, and K. M. Lim, "RoBERTa-LSTM: A Hybrid Model for Sentiment Analysis With Transformer and Recurrent Neural Network," *IEEE Access*, vol. 10, pp. 21517–21525, 2022, doi: 10.1109/ACCESS.2022.3152828.

[5] N. Nurzaenab *et al.*, "EVALUASI PERFORMA METODE LONG SHORT TERM MEMORY (LSTM) DAN RECURRENT NEURAL NETWORK (RNN) PADA ANALISIS SENTIMEN KOMENTAR PENGGUNA APLIKASI KITALULUS," *JTRISTE*, vol. 12, pp. 86–92, Sep. 2025, doi: 10.55645/jtriste.v12i1.614.

[6] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Comput.*, vol. 9, pp. 1735–1780, Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.

[7] B. Wilie *et al.*, "IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding," in *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing, AACL-IJCNLP 2020*, Association for Computational Linguistics (ACL), 2020, pp. 843–857. doi: 10.18653/v1/2020.aacl-main.85.

[8] S. Aulia Natasya, N. Muzahra, A. Baihaqi, R. Lesmana, and Nurbaiti Nurbaiti, "Persepsi Konsumen terhadap Consumer Review pada Platform E-Commerce Shopee," *Jurnal Nuansa : Publikasi Ilmu Manajemen dan Ekonomi Syariah*, vol. 3, pp. 196–207, Dec. 2025, doi: 10.61132/nuansa.v3i4.2406.

[9] A. Ahmad, Aswadi Ramli, and H. Hajerah, "Penggunaan Bahasa Gaul di Kalangan Remaja terhadap Kelestarian Bahasa Indonesia di Era Digital," *Jurnal Onoma: Pendidikan, Bahasa, dan Sastra*, vol. 11, pp. 980–990, Feb. 2025, doi: 10.30605/onoma.v11i1.5018.

[10] Pande sindu, Agus Aan Jiwa Permana, and I Nyoman Saputra Wahyu Wijaya, "IDENTIFIKASI DAN NORMALISASI TEKS SLANG DENGAN FASTTEXT PADA TWITTER DALAM BAHASA INDONESIA," *Jurnal Pendidikan Teknologi dan Kejuruan*, vol. 21, pp. 33–44, Jan. 2024, doi: 10.23887/jptkundiksha.v21i1.66381.

[11] D. A. Kristiyanti, S. A. Sanjaya, V. C. Tjokro, and J. Suhali, "Dealing imbalance dataset problem in sentiment analysis of recession in Indonesia," *IAES International Journal of Artificial Intelligence*, vol. 13, pp. 2058–2070,

Jun. 2024, doi: 10.11591/ijai.v13.i2.pp2060-2072.

[12] S. Riyanto, I. S. Sitanggang, T. Djatna, and T. D. Atikah, "Comparative Analysis using Various Performance Metrics in Imbalanced Data for Multi-class Text Classification," *International Journal of Advanced Computer Science and Applications*, vol. 14, pp. 1082–1090, 2023, doi: 10.14569/IJACSA.2023.01406116.

[13] M. Idris, A. Rifai, and K. D. Tania, "Sentiment Analysis of Tokopedia App Reviews using Machine Learning and Word Embeddings," *sinkron*, vol. 9, pp. 210–219, Jan. 2025, doi: 10.33395/sinkron.v9i1.14278.

[14] F. M. Bianchi, E. Maiorino, M. C. Kampffmeyer, A. Rizzi, and R. Jenssen, "An overview and comparative analysis of Recurrent Neural Networks for Short Term Load Forecasting," Jul. 2018, doi: 10.1007/978-3-319-70338-1.

[15] I. Muis, "Pemetaan Literatur tentang Layanan Prima di Perguruan Tinggi: Analisis Tematik Berbasis PRISMA," *Jurnal Siber Multi Disiplin*, vol. 3, pp. 12–24, May 2025, doi: 10.38035/jsmd.v3i1.418.

[16] D. Ariffadillah, R. R. P. C. -, S. M. R. -, R. L. Pratiwi, and R. I. Handayani, "KOMPARASI METODE SVM DAN BILSTM PADA KLASIFIKASI SENTIMEN APLIKASI PLAY STORE DENGAN TEKNIK HYBRID IMBALANCE HANDLING," *Jurnal Informatika dan Teknik Elektro Terapan*, vol. 14, Jan. 2026, doi: 10.23960/jitet.v14i1.8313.

[17] M. J. Page *et al.*, "PRISMA 2020 explanation and elaboration: Updated guidance and exemplars for reporting systematic reviews," Mar. 2021, *BMJ Publishing Group*. doi: 10.1136/bmj.n160.

[18] B. Kitchenham, O. Pearl Brereton, D. Budgen, M. Turner, J. Bailey, and S. Linkman, "Systematic literature reviews in software engineering - A systematic literature review," Jan. 2009. doi: 10.1016/j.infsof.2008.09.009.

[19] D. Moher *et al.*, "Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement," Jul. 2009, *Public Library of Science*. doi: 10.1371/journal.pmed.1000097.

[20] J. Cohen, "A Coefficient of Agreement for Nominal Scales," *Educ. Psychol. Meas.*, vol. 20, pp. 37–46, 1960, doi: 10.1177/001316446002000104.

[21] B. Hakim, "Analisa Sentimen Data Text Preprocessing Pada Data Mining Dengan Menggunakan Machine Learning," *JBASE - Journal of Business and Audit Information Systems*, vol. 4, no. 2, Aug. 2021, doi: 10.30813/jbase.v4i2.3000.

[22] M. I. Raif, N. N. Hidayati, and T. Matulatan, "Otomatisasi Pendeteksi Kata Baku Dan Tidak Baku Pada Data Twitter Berbasis KBBI," *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 11, pp. 337–348, Apr. 2024, doi: 10.25126/jtiik.20241127404.

[23] B. A. Haryono and A. Prasetya, "Identifikasi Kuantitas Berbasis Rule Pada Masalah Text-to-SQL," *Jurnal Borneo Informatika dan Teknik Komputer*, vol. 5, pp. 26–39, Sep. 2025, doi: 10.35334/jbit.v5i1.6819.

[24] M. I. Raif, N. N. Hidayati, and T. Matulatan, "Otomatisasi Pendeteksi Kata Baku Dan Tidak Baku Pada Data Twitter Berbasis KBBI," *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 11, pp. 337–348, Apr. 2024, doi: 10.25126/jtiik.20241127404.

[25] M. T. Ramadhan, C. Sobarna, and A. S. Afsari, "Code Mixing of Slang and Sundanese on TikTok," *SUAR BETANG*, vol. 18, no. 2, pp. 265–276, Dec. 2023, doi: 10.26499/surbet.v18i2.13751.

[26] M. F. R. Khairul and R. S. Perdana, "Arsitektur Sistem Percakapan Otomatis Berbahasa Indonesia dengan Normalisasi Bahasa Informal Menjadi Baku," *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 11, pp. 1009–1016, Oct. 2024, doi: 10.25126/jtiik.2024117984.

[27] Asnawiyah and R. Eka Putra, "Perbandingan Algoritma LSTM dan BiLSTM Untuk Analisis Sentimen Multi-Class Media Sosial Twitter," *Journal of Informatics and Computer Science (JINACS)*, vol. 6, pp. 778–786, Jan. 2025, doi: 10.26740/jinacs.v6n03.p778-786.

[28] H. Setiawan and D. Ariatmanto, "Analisis Perbandingan Algoritma Machine Learning Dan Deep Learning Untuk Sentimen Analisis Teks Umpan Balik Tentang Evaluasi Pengajaran Dosen," *JSAI (Journal Scientific and Applied Informatics)*, vol. 7, pp. 379–385, Jun. 2024, doi: 10.36085/jsai.v7i2.6572.

[29] A. R. Isnain, H. Sulistiani, B. M. Hurohman, A. Nurkholis, and S. Styawati, "Analisis Perbandingan Algoritma LSTM dan Naive Bayes untuk Analisis Sentimen," *Jurnal Edukasi dan Penelitian Informatika (JEPIN)*, vol. 8, p. 299, Aug. 2022, doi: 10.26418/jp.v8i2.54704.

[30] Yuyun, A. D. Latief, T. Sampurno, Hazriani, A. O. Arisha, and Mushaf, "Next Sentence Prediction: The Impact of Preprocessing Techniques in Deep Learning," in *Proceedings - 2023 10th International Conference on Computer, Control, Informatics and its Applications: Exploring the Power of Data: Leveraging Information to Drive Digital Innovation, IC3INA 2023*,

Institute of Electrical and Electronics Engineers Inc., 2023, pp. 274–278. doi: 10.1109/IC3INA60834.2023.10285805.

[31] D. Purnamasari, A. B. Aji, S. Madenda, I. M. Wiryana, and S. Harmanto, "SENTIMENT ANALYSIS METHODS FOR CUSTOMER REVIEW OF INDONESIA E-COMMERCE," *International Journal of Innovative Computing, Information and Control*, vol. 20, no. 1, pp. 47–60, Feb. 2024, doi: 10.24507/ijicic.20.01.47.

[32] A. Romadhony, S. Al Faraby, R. Rismala, U. N. Wisesti, and A. Arifianto, "Sentiment Analysis on a Large Indonesian Product Review Dataset," *Journal of Information Systems Engineering and Business Intelligence*, vol. 10, pp. 167–178, 2024, doi: 10.20473/jisebi.10.1.167-178.

[33] N. G. Ramadhan, "Comparative Analysis of ADASYN-SVM and SMOTE-SVM Methods on the Detection of Type 2 Diabetes Mellitus," *Scientific Journal of Informatics*, vol. 8, pp. 276–282, Nov. 2021, doi: 10.15294/sji.v8i2.32484.

[34] T. I. Z. M. Putra, S. Suprapto, and A. F. Bukhori, "Model Klasifikasi Berbasis Multiclass Classification dengan Kombinasi Indobert Embedding dan Long Short-Term Memory untuk Tweet Berbahasa Indonesia," *Jurnal Ilmu Siber dan Teknologi Digital*, vol. 1, pp. 1–28, Nov. 2022, doi: 10.35912/jisted.v1i1.1509.

[35] M. I. Alfarizi, L. Syafaah, and M. Lestandy, "Emotional Text Classification Using TF-IDF (Term Frequency-Inverse Document Frequency) And LSTM (Long Short-Term Memory)," *JUITA : Jurnal Informatika*, vol. 10, p. 225, Nov. 2022, doi: 10.30595/juita.v10i2.13262.

[36] H. Hairani, T. Widiyaningtyas, and D. D. Prasetya, "Addressing Class Imbalance of Health Data: a Systematic Literature Review on Modified Synthetic Minority Oversampling Technique (SMOTE) Strategies," *International Journal on Informatics Visualization*, vol. 8, pp. 1310–1318, 2024, doi: 10.62527/joiv.8.3.2283.

[37] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002, doi: 10.1613/jair.953.

[38] H. He, Y. Bai, E. A. Garcia, and S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," in *Proceedings of the International Joint Conference on Neural Networks*, 2008, pp. 1322–1328. doi: 10.1109/IJCNN.2008.4633969.

[39] Cosmas Haryawan and Yosef Muria Kusuma Ardhana, "ANALISA PERBANDINGAN TEKNIK OVERSAMPLING SMOTE PADA IMBALANCED DATA," *Jurnal Informatika dan Rekayasa Elektronik*, vol. 6, pp. 73–78, Apr. 2023, doi: 10.36595/jire.v6i1.834.

[40] R. Waluyo and A. S. Munir, "Optimasi Prediksi Kematian pada Gagal Jantung Analisis Perbandingan Algoritma Pembelajaran Ensemble dan Teknik Penyeimbangan Data pada Dataset," *Jurnal Sistem dan Teknologi Informasi (JustIN)*, vol. 12, p. 365, Apr. 2024, doi: 10.26418/justin.v12i2.75158.

[41] I. G. Ayu Nandia Lestari, D. G. Hendra Divayana, and K. Y. Ernada Aryanto, "A Concentration Selection In Study Programs Using SMOTE Techniques With Ensemble Learning Algorithms," in *2023 5th International Conference on Cybernetics and Intelligent Systems, ICORIS 2023*, Institute of Electrical and Electronics Engineers Inc., 2023. doi: 10.1109/ICORIS60118.2023.10352192.

[42] P. Branco, L. Torgo, and R. P. Ribeiro, "A survey of predictive modeling on imbalanced domains," Jun. 2017, *Association for Computing Machinery*. doi: 10.1145/2907070.

[43] M. A. Hermawan, A. Faqih, and G. Dwilestari, "IMPLEMENTASI AKURASI MODEL NAIVE BAYES MENGGUNAKAN SMOTE DALAM ANALISIS SENTIMEN PENGGUNA APLIKASI BRIMO," *Jurnal Informatika dan Teknik Elektro Terapan*, vol. 13, Jan. 2025, doi: 10.23960/jitet.v13i1.5748.

[44] D. Ariffadillah, R. R. P. C. -, S. M. R. -, R. L. Pratiwi, and R. I. Handayani, "KOMPARASI METODE SVM DAN BILSTM PADA KLASIFIKASI SENTIMEN APLIKASI PLAY STORE DENGAN TEKNIK HYBRID IMBALANCE HANDLING," *Jurnal Informatika dan Teknik Elektro Terapan*, vol. 14, Jan. 2026, doi: 10.23960/jitet.v14i1.8313.

[45] J. Heaton, "Ian Goodfellow, Yoshua Bengio, and Aaron Courville: Deep learning," *Genet. Program. Evolvable Mach.*, vol. 19, pp. 305–307, Jun. 2018, doi: 10.1007/s10710-017-9314-z.

[46] C. Sintiya, G. H. Hutagaol, D. Bate`e, and S. Irviantina, "Evaluasi Teknik Resampling untuk Class Balancing dalam Analisis Sentimen Kesehatan Mental Berbasis Bi-LSTM," *Jurnal Sifo Mikroskil*, vol. 26, Oct. 2025, doi: 10.55601/jsm.v26i2.1799.

[47] J. P. A. Ioannidis, "Why most published research findings are false," in *Getting to Good: Research Integrity in the Biomedical Sciences*, Springer International Publishing,

2018, pp. 2–8. doi: 10.1371/journal.pmed.0020124.
[48] I. Maharani, I. Alifiar, and Y. P. Sukmawan, “Analisis Bibliometrik Terkait Tren Teratogenik,” *Pharmacy Genius*, vol. 4, pp. 127–134, Oct. 2025, doi: 10.56359/pharmgen.v4i3.912.
[49] I. B. Saputro and T. J. Ai, “PENGGUNAAN ANALISIS SENTIMEN UNTUK PERANCANGAN PRODUK: SEBUAH TINJAUAN PUSTAKA,” *Jurnal Rekavasi*, vol. 12, pp. 14–22, Jun. 2024, doi: 10.34151/rekavasi.v12i1.4660.
[50] M. Harsanto and E. Sudarmilah, “TINJAUAN LITERATUR ANALISIS SENTIMEN PRODUK E-COMMERCE: DATASET, PENDEKATAN, METODE, DAN PERFORMA,” *JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)*, vol. 10, pp. 2290–2303, Aug. 2025, doi: 10.29100/jipi.v10i3.8217.