INTI NUSA MANDIRI
DOI: https://doi.org/10.33480/inti.v21i1.8587

VOL. 21. NO. 1 AGUSTUS 2026
P-ISSN: 0216-6933 | E-ISSN: 2685-807X
Diterbitkan Oleh: LPPM Universitas Nusa Mandiri

DETEKSI SPOILER PADA ULASAN BUKU BERBAHASA INDONESIA MENGGUNAKAN PENDEKATAN MACHINE LEARNING DAN DEEP LEARNING

Natasya Agustine Sadhi[1]; Hannah Larissa Halim[2]; Jessica Winola[3]; Viny Christanti Mawardi[4*]

Program Studi Teknik Informatika[1,2,3,4]
Universitas Tarumanagara, Jakarta, Indonesia [1,2,3,4]
www.untar.ac.id [1,2,3,4]
natasya.535240001@stu.untar.ac.[1], hannah.535240015@stu.untar.ac.id[2],
jessica.535240026@stu.untar.ac.id[3], vinym@fti.untar.ac.id[4*]

(*) Corresponding Author

Abstract— *Spoiler detection in book reviews is a challenging text classification task because spoilers are not always identifiable from specific words but depend heavily on the narrative context in which information is revealed. This study compares six text classification models for detecting spoilers in Indonesian-language book reviews from Goodreads, namely Support Vector Machine (SVM),* Random Forest*, XGBoost, Bidirectional Long Short-Term Memory (BiLSTM), IndoBERT, and XLM-R (RoBERTa). Data were collected through Selenium-based web scraping and GraphQL API from 40 mystery and thriller book titles, resulting in 11,259 reviews with a class imbalance ratio of 1:10.4. All models were evaluated using AUC-ROC, PR-AUC, spoiler F1-score, and spoiler recall as primary metrics, with decision thresholds optimized through each model's validation set. XLM-R achieved the best overall performance with an AUC-ROC of 0.7017, a PR-AUC of 0.2263, and a spoiler F1-score of 0.2903, followed by IndoBERT, SVM, Random Forest, XGBoost, and BiLSTM. Overall, transformer-based models outperformed the traditional machine learning models and BiLSTM across most evaluation metrics. The results also indicate that each model exhibits different precision-recall characteristics, suggesting that model performance should not be evaluated using a single metric alone. These findings can serve as an initial reference for future research on Indonesian-language spoiler detection and support the development of automated spoiler detection systems for content moderation on digital book review platforms.*

Keywords*: BiLSTM, Class Imbalance, IndoBERT, Machine Learning, Spoiler Detection.*

Abstrak— Deteksi *spoiler* pada ulasan buku merupakan tugas klasifikasi teks yang menantang karena *spoiler* tidak selalu dapat dikenali dari kata-kata tertentu, melainkan bergantung pada konteks narasi tempat informasi tersebut diungkapkan. Penelitian ini membandingkan enam model klasifikasi teks untuk mendeteksi *spoiler* pada ulasan buku berbahasa Indonesia dari Goodreads, yaitu *Support Vector Machine* (SVM), Random Forest, XGBoost, *Bidirectional Long Short-Term Memory* (BiLSTM), IndoBERT, dan XLM-R (RoBERTa). Data dikumpulkan melalui *web scraping* berbasis Selenium dan GraphQL API dari 40 judul buku bergenre misteri dan *thriller*, menghasilkan 11.259 ulasan dengan rasio ketidakseimbangan kelas sebesar 1:10,4. Seluruh model dievaluasi menggunakan AUC-ROC, PR-AUC, F1 *spoiler*, dan *recall spoiler* sebagai metrik utama, dengan *threshold* yang dioptimalkan melalui *validation set* masing-masing model. XLM-R memperoleh performa terbaik dengan AUC-ROC 0,7017, PR-AUC 0,2263, dan F1-score spoiler 0,2903, diikuti oleh IndoBERT, SVM, Random Forest, XGBoost, dan BiLSTM. Secara umum, model berbasis transformer menghasilkan performa yang lebih baik dibandingkan model machine learning tradisional dan BiLSTM pada sebagian besar metrik evaluasi. Hasil penelitian juga menunjukkan bahwa setiap model memiliki karakteristik precision dan recall yang berbeda sehingga evaluasi performa tidak dapat didasarkan pada satu metrik saja. Hasil ini dapat menjadi rujukan awal bagi penelitian deteksi spoiler berbahasa Indonesia selanjutnya serta mendukung pengembangan sistem deteksi spoiler otomatis untuk membantu moderasi konten pada platform ulasan buku digital.

Kata kunci: *BiLSTM, Ketidakseimbangan Kelas, IndoBERT, Machine Learning, Deteksi Spoiler.*

VOL. 21. NO. 1 AGUSTUS 2026
P-ISSN: 0216-6933 | E-ISSN: 2685-807X
Diterbitkan Oleh: LPPM Universitas Nusa Mandiri
INTI NUSA MANDIRI
DOI: https://doi.org/10.33480/inti.v21i1.8587

PENDAHULUAN

Pertumbuhan masif platform ulasan buku digital seperti Goodreads telah memicu pergeseran paradigma dalam konsumsi literatur melalui fenomena *participatory reading.* Dalam ekosistem ini, pembaca tidak lagi sekadar menjadi konsumen pasif, melainkan aktor aktif yang memproduksi dan mendistribusikan opini naratif secara daring. Namun, volume konten buatan pengguna (*user-generated content*) yang sangat besar ini membawa tantangan signifikan berupa penyebaran *spoiler*. Secara konseptual, *spoiler* didefinisikan sebagai pengungkapan detail krusial dari sebuah karya media, seperti resolusi konflik atau *plot twist* yang tidak terduga sebelum pembaca menyelesaikan karya tersebut.

Keberadaan *spoiler* menjadi isu kritis karena secara psikologis dapat mengurangi antisipasi kognitif dan kepuasan estetis pembaca secara fundamental (Wróblewska et al., 2021). Urgensi deteksi *spoiler* didukung oleh temuan yang menyatakan bahwa antisipasi merupakan komponen vital dalam kenikmatan naratif, sehingga pengungkapan informasi secara prematur berdampak langsung pada penurunan keterlibatan emosional pembaca. Masalah ini menjadi sangat relevan pada genre buku misteri, *thriller,* dan horor yang secara inheren bergantung pada elemen kejutan dan ketidakpastian plot untuk membangun ketegangan.

Terbukti dari data bahwa genre *mystery*, *thriller*, dan *crime* memiliki persentase ulasan mengandung spoiler tertinggi dibandingkan genre lainnya (Wróblewska et al., 2021). Meskipun platform digital telah menyediakan fitur pelaporan berbasis pengguna (*crowdsourcing*), mekanisme tersebut terbukti tidak memadai untuk menangani laju pertumbuhan volume ulasan yang eksponensial dan sering kali menunjukkan inkonsistensi subjektivitas yang tinggi. Hal ini mendorong perlunya pengembangan sistem deteksi spoiler otomatis berbasis *Natural Language Processing* (NLP) yang mampu melakukan moderasi konten secara presisi.

Penelitian terbaru dalam ranah NLP menunjukkan transisi dari model perhatian hierarkis tradisional menuju penggunaan *Graph Neural Network* (SDGNN) yang mampu menangkap relasi dependensi sintaktis secara kompleks (Chang et al., 2021). Selain itu, tinjauan terhadap berbagai metode deteksi spoiler menegaskan bahwa arsitektur berbasis *Long Short-Term Memory* (LSTM) dan *Transformer* kini menjadi standar dalam membedakan struktur informasi sensitif pada ulasan buku, dengan pendekatan *fine-tuning Transformer* seperti BERT dan ELECTRA menunjukkan performa terbaik pada dataset spoiler skala besar (Wróblewska et al., 2021). Bahkan, penggunaan model sekuensial tanpa fitur buatan tangan (*handcrafted*) telah terbukti mampu melampaui performa model-model dasar sebelumnya (Bao et al., 2021).

Namun, mayoritas studi tersebut masih berfokus eksklusif pada korpus berbahasa Inggris, sehingga diperlukan upaya pengembangan model yang mampu menangani sumber daya bahasa lokal. Bahasa Indonesia memiliki karakteristik linguistik unik yang memerlukan perlakuan khusus, mencakup sistem morfologi atau afiksasi yang kompleks serta penggunaan bahasa informal yang dominan pada ulasan digital (Lubis et al., 2023). Pentingnya penggunaan model yang telah di-*pretrain* secara khusus pada korpus lokal menjadi sangat krusial untuk menangkap nuansa semantik bahasa Indonesia yang tidak terwakili dalam model multilingual (Koto et al., 2020).

Terdapat pula kata-kata transisi spesifik yang secara leksikal berfungsi sebagai indikator kuat kemunculan spoiler dalam narasi bahasa Indonesia, namun sering kali tereduksi dalam proses pengolahan teks standar jika tidak ditangani dengan tepat. Kesenjangan penelitian ini diperparah oleh ketiadaan studi komparatif sistematis yang mengevaluasi performa antara algoritma *Machine Learning* tradisional dan *Deep Learning* khusus untuk deteksi *spoiler* berbahasa Indonesia.

Justifikasi perbandingan ini krusial karena model ML tradisional seperti SVM dikenal efisien dalam menangani ruang fitur yang terbatas (Jamshidian, 2023). Sementara itu, model *Deep Learning* seperti BiLSTM menawarkan keunggulan dalam menangkap informasi kontekstual dan dependensi sekuensial dari dua arah, sehingga mampu memahami hubungan antar kata secara lebih komprehensif dalam teks berbahasa alami (Lestari, 2024). Berdasarkan latar belakang tersebut, penelitian ini bertujuan untuk mengevaluasi ketahanan (*robustness*) algoritma *machine learning* tradisional dan *deep learning* dalam menangani masalah ketidakseimbangan kelas yang ekstrem (1:10,4) pada tugas deteksi *spoiler* ulasan buku berbahasa Indonesia.

Penggunaan model IndoBERT secara spesifik dipilih karena keunggulannya sebagai standar *benchmark* dalam tugas klasifikasi teks bahasa Indonesia yang mampu menangkap nuansa bahasa informal dengan lebih akurat (Koto et al., 2020; Zevana & Riana, 2024). Selain itu, model XLM-RoBERTa (XLM-R) juga dievaluasi sebagai

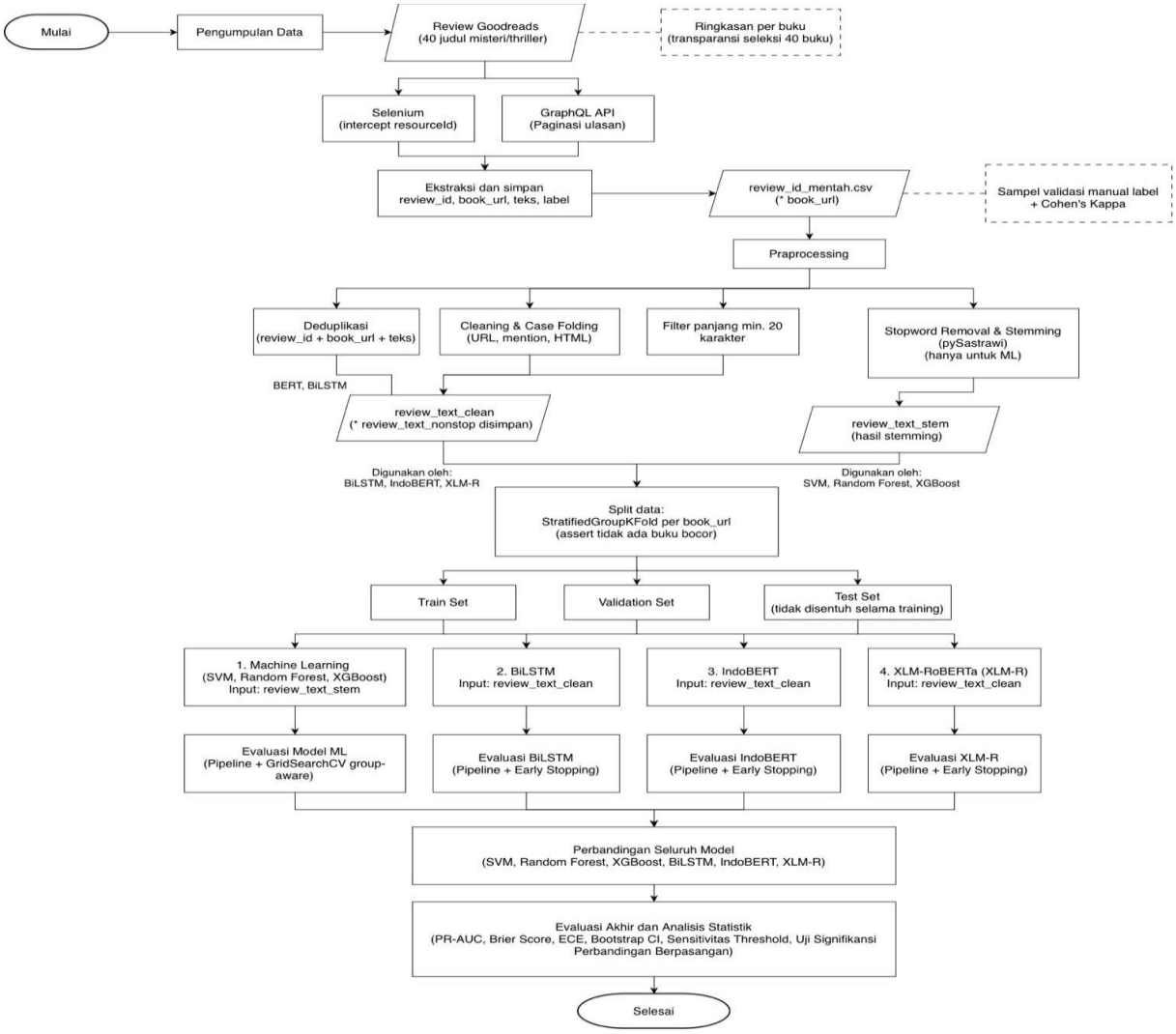
INTI NUSA MANDIRI
DOI: https://doi.org/10.33480/inti.v21i1.8587

VOL. 21. NO. 1 AGUSTUS 2026
P-ISSN: 0216-6933 | E-ISSN: 2685-807X
Diterbitkan Oleh: LPPM Universitas Nusa Mandiri

representasi Transformer multibahasa untuk menguji efektivitas model lintas bahasa yang dilatih pada skala besar pada tugas klasifikasi ulasan lokal. Melalui penelitian ini, diharapkan dapat dihasilkan referensi empiris mengenai sejauh mana kompleksitas arsitektur model mampu mengatasi bias kelas mayoritas pada sumber daya teks lokal, sehingga memberikan kontribusi spesifik bagi penanganan data yang tidak seimbang (imbalanced data) dalam moderasi konten literasi digital.

BAHAN DAN METODE

Penelitian ini menggunakan data ulasan buku dari Goodreads yang menyediakan fitur penandaan spoiler sehingga label spoiler dan non-spoiler dapat diperoleh secara langsung. Penelitian ini bertujuan membandingkan kinerja beberapa algoritma klasifikasi teks untuk mendeteksi *spoiler* pada ulasan buku berbahasa Indonesia. Tahapan metodologi penelitian ditunjukkan pada Gambar 1.



Sumber: (Hasil Penelitian, 2026)

Gambar 1. Tahapan Metodologi

Alur penelitian pada Gambar 1 mencakup enam tahapan: pengumpulan data (disertai validasi label manual menggunakan Cohen's Kappa), praproses data, pembagian data (StratifiedGroupKFold berbasis book_url untuk mencegah kebocoran antar buku), pemodelan, perbandingan model, dan evaluasi akhir, yang dijelaskan pada subbab berikut.

Pengumpulan Data

Data dikumpulkan dari Goodreads menggunakan web scraping berbasis Selenium dan GraphQL API pada Mei 2026, dengan label spoiler diperoleh dari penandaan pengguna Goodreads. Untuk menguji keandalan label spoiler bawaan Goodreads, dilakukan validasi manual terhadap 300 sampel ulasan (proporsional antara kelas spoiler

dan non-spoiler) oleh dua annotator independen yang tidak mengetahui label asli saat proses anotasi berlangsung. Kesepakatan antar kedua annotator tergolong sangat tinggi (κ = 0,963; *almost perfect*), menunjukkan bahwa proses anotasi manual dilakukan secara konsisten antar-penilai. Namun, kesepakatan antara label bawaan Goodreads dan label hasil anotasi manual hanya tergolong *fair* (κ= 0,367 terhadap annotator 1 dan 0,340 terhadap annotator 2, dengan tingkat kesepakatan mentah masing-masing sebesar 68,3% dan 67,0%).

Temuan ini mengindikasikan bahwa label bawaan Goodreads memiliki keterbatasan validitas yang cukup berarti dibandingkan penilaian manusia, sehingga dibahas lebih lanjut pada bagian Keterbatasan Penelitian. Pemilihan 40 judul buku dilakukan melalui purposive sampling dengan kriteria ketersediaan edisi berbahasa Indonesia dan volume ulasan memadai, mencakup kombinasi penulis lokal dan terjemahan genre misteri dan thriller (Brown, 2025); jumlah ulasan per buku bervariasi 1–1.210 (median 225).

Setelah proses deduplikasi, filter bahasa Indonesia, filter panjang minimum 20 karakter, dan penghapusan baris dengan nilai kosong pada kolom utama, dari 11.336 ulasan mentah diperoleh 11.259 ulasan akhir (988 spoiler, 8,78%; 10.271 non-spoiler, 91,22%), rasio ketidakseimbangan 1:10,4 (total data terhapus: 77 ulasan atau 0,7%). Rincian lengkap mengenai daftar 40 judul buku beserta jumlah ulasan yang berhasil ditarik untuk setiap buku disajikan pada Tabel 1.

Tabel 1. Distribusi Jumlah Ulasan dan Persentase Ulasan Spoiler per Buku

No	Judul Buku	Jumlah Ulasan	Persentase Spoiler
1	Teka-Teki Rumah Aneh	1.201	10,03%
2	Malice	859	10,94%
3	Holy Mother	820	11,34%
4	Girls in the Dark	706	7,79%
5	Second Sister	610	9,18%
6	Rumah Lebah	535	12,34%
7	The Tokyo Zodiac Murders	467	8,57%
8	Confessions	441	6,58%
9	Pasien	399	8,02%
10	The Dead Returns	369	5,42%
11	Black Showman dan Pembunuhan di Kota Tak Bernama	354	6,50%
12	Dua Dini Hari	326	8,28%
13	Misteri Patung Garam	284	4,23%
14	Penance	279	11,11%
15	Tragedi Pedang Keadilan	267	19,85%
16	Angsa dan Kelelawar	264	8,71%
17	A Midsummer's Equation	255	9,02%
18	The Good Son	252	12,70%
19	24 Jam Bersama Gaspar	242	3,72%
20	Murder in the Crooked House	225	8,00%
21	Burning Heat	221	15,38%
22	Rahasia Meede	217	2,30%
23	Danur	204	2,45%
24	About a Place in the Kinki Region	164	6,10%
25	Omen	159	8,81%
26	The Borrowed	151	5,30%
27	Obsesi	135	0,00%
28	Playing Victim	97	6,19%
29	Pengurus MOS Harus Mati	93	4,30%
30	Ve	92	9,78%
31	Tujuh Lukisan Horor	82	3,66%
32	Malam Karnaval Berdarah	80	3,75%
33	Misteri Organisasi Rahasia The Judges	79	1,27%
34	Permainan Maut	75	1,33%
35	Teror	71	8,45%
36	Target Terakhir	69	15,94%
37	Kutukan Hantu Opera	64	6,25%
38	Sang Pengkhianat	47	17,02%
39	Girls in the Dark	2	0,00%
40	Guilt	1	0,00%

Sumber: (Hasil Penelitian, 2026)

Praproses Data

Tahap ini meliputi deduplikasi berdasarkan review_id, book_url, dan teks ulasan, deteksi bahasa menggunakan langdetect (Wyawhare, 2023), pembersihan teks (URL, mention, hashtag, HTML) dan case folding, serta penghapusan ulasan dengan panjang kurang dari 20 karakter. Stopword removal (mempertahankan indikator spoiler) dan stemming menggunakan PySastrawi (Suhaeni et al., 2025) diterapkan khusus untuk model machine learning tradisional. Dataset akhir disiapkan dalam dua bentuk: teks bersih tanpa stemming untuk BiLSTM/IndoBERT/XLM-R, dan teks hasil stemming untuk SVM/Random Forest/XGBoost, karena perbedaan kebutuhan representasi teks (Lubis et al., 2023).

Pemodelan

Penelitian ini menggunakan enam algoritma dari tiga pendekatan: machine learning tradisional (SVM, Random Forest, XGBoost), deep learning (BiLSTM), dan Transformer (IndoBERT, XLM-R). SVM dipilih karena efektif pada fitur berdimensi tinggi seperti TF-IDF (Al-Habib et al., 2023; Jamshidian, 2023) Random Forest karena tahan overfitting melalui ensemble, dan XGBoost karena kemampuan gradient boosting-nya

VOL. 21. NO. 1 AGUSTUS 2026
P-ISSN: 0216-6933 | E-ISSN: 2685-807X
Diterbitkan Oleh: LPPM Universitas Nusa Mandiri

memperbaiki kesalahan secara sekuensial. BiLSTM dipilih karena memproses teks dua arah untuk representasi kontekstual yang lebih kaya (Lestari, 2024), sedangkan IndoBERT karena self-attention-nya menangkap relasi antar token global dengan bobot pre-training korpus bahasa Indonesia (Koto et al., 2020). XLM-R ditambahkan sebagai pembanding Transformer multilingual (Conneau et al., 2020) untuk mengevaluasi pengaruh korpus pre-training lintas bahasa.

Representasi fitur untuk model machine learning tradisional dibentuk menggunakan TF-IDF (Zhang, 2025) sebagaimana ditunjukkan pada Persamaan (1), sedangkan BiLSTM menggunakan pretrained FastText bahasa Indonesia berdimensi 300 (Lestari, 2024) dan IndoBERT maupun XLM-R menggunakan tokenizer bawaan masing-masing dengan panjang maksimum 256 token (Zevana & Riana, 2024). Untuk mengevaluasi dampak pemotongan teks akibat batas panjang token ini, dilakukan analisis performa terpisah antara ulasan yang terpotong dan tidak terpotong pada data uji.

$$TF - IDF(t,d) = TF(t,d) \times IDF(t) \quad (1)$$

Arsitektur BiLSTM terdiri atas embedding layer FastText, lapisan Bidirectional LSTM, Dropout, Dense dengan regularisasi L2, dan output Dense berfungsi aktivasi sigmoid, dilatih dalam dua fase gradual unfreezing (embedding dibekukan lalu dibuka dengan learning rate lebih kecil untuk fine-tuning) (Karakaya & Kilimci, 2024). IndoBERT dan XLM-R di-fine-tune masing-masing dari checkpoint indobert-base-p2 dan xlm-roberta-base dengan konfigurasi identik (Tabel 3). Ketidakseimbangan kelas ditangani melalui weighted cross-entropy loss pada kedua model Transformer serta parameter class_weight pada model machine learning tradisional.

Untuk mencegah kebocoran data (*data leakage*) antar-buku secara konsisten, pembagian data dilakukan dengan skema dua tahap berbasis kelompok (*group-preserved*) menggunakan variabel *book_url* sebagai kunci kelompok. Tahap pertama menggunakan prinsip StratifiedGroupKFold untuk memisahkan data uji sebesar 20% (2.252 ulasan) dari keseluruhan dataset utuh (11.260 ulasan). Tahap kedua membagi sisa data (80% atau 9.008 ulasan) menjadi data latih (85% dari sisa data atau 7.657 ulasan) dan data validasi (15% dari sisa data atau 1.351 ulasan) dengan tetap mengunci *book_url*. Prosedur ini menjamin secara absolut bahwa ulasan dari suatu buku yang sama hanya akan berada di salah satu himpunan data saja (latih, validasi, atau uji) sehingga tidak ada tumpang tindih informasi. Evaluasi menggunakan satu kali pembagian data (*single-split*) dari lipatan tersebut, bukan *cross-validation* berulang karena tingginya biaya komputasi *fine-tuning* Transformer. Perbedaan prosedur tuning dan skema *single-split* ini diakui sebagai keterbatasan metodologis (lihat bagian Keterbatasan). Threshold klasifikasi ditentukan dari *precision-recall curve* pada *validation set* (F1-score maksimum), dengan probabilitas SVM dikalibrasi via *Platt scaling*; sensitivitas threshold dianalisis pada rentang 0,05–0,95 (interval 0,01). Seluruh konfigurasi hyperparameter disajikan pada Tabel 2.

Tabel 2. Konfigurasi *Hyperparameter* Seluruh Model

Model	*Hyperparameter*	Nilai
SVM	C	0,01
	TF-IDF *max_features*	30000
	TF-IDF *ngram_range*	(1,1)
Random Forest	*n_estimators*	200
	max_depth	None
	max_features	sqrt
	min_samples_leaf	10
	TF-IDF *max_features*	50000
	class_weight	balanced
	TF-IDF *ngram_range*	(1,2)
XGBoost	*n_estimators*	100
	max_depth	3
	learning_rate	0,05
	subsample	0,7
	colsample_bytree	0,7
	scale_pos_weight	5
	TF-IDF *max_features*	50000
	class_weight	balanced
	TF-IDF *ngram_range*	(1,2)
BiLSTM	LSTM *units* (per arah)	64
	Dropout	0,5
	Dense units	32
	L2 regularization	0,01
	Learning rate fase 1	1×10^{-3}
	Learning rate fase 2	1×10^{-4}
	Batch size	64
	Early stopping patience	5
IndoBERT	*Learning rate*	5×10^{-6}
	Weight decay	0,1
	Warmup	1 *Epoch*
	Max token length	256
	Batch size	16
	Early stopping patience	2
XLM-R	*Learning rate*	5×10^{-6}
	Weight decay	0,1
	Warmup	1 *Epoch*
	Max token length	256
	Batch size	16
	Early stopping patience	2

Sumber: (Hasil Penelitian, 2026)

Pembagian data melalui skema dua tahap berbasis kelompok (*group-preserved*) tersebut dilakukan menggunakan StratifiedGroupKFold sebanyak dua kali. Tahap pertama memisahkan

VOL. 21. NO. 1 AGUSTUS 2026
P-ISSN: 0216-6933 | E-ISSN: 2685-807X
Diterbitkan Oleh: LPPM Universitas Nusa Mandiri
INTI NUSA MANDIRI
DOI: https://doi.org/10.33480/inti.v21i1.8587

data uji dengan StratifiedGroupKFold lima lipatan (*outer fold*, 5 *splits*), di mana satu lipatan diambil sebagai data uji dan sisanya digunakan sebagai data latih-validasi gabungan. Tahap kedua memisahkan data validasi dari data latih-validasi gabungan tersebut menggunakan StratifiedGroupKFold enam lipatan (*inner fold*, 6 *splits*), dengan satu lipatan diambil sebagai data validasi. Skema ini menghasilkan 6.546 sampel latih dari 26 judul buku (8,66% spoiler), 982 sampel validasi dari 5 judul buku (9,57% spoiler), dan 3.731 sampel uji dari 9 judul buku (8,76% spoiler).

Verifikasi eksplisit dilakukan untuk memastikan tidak ada judul buku (book_url) yang tumpang tindih antara ketiga partisi data. Proses pembagian dijalankan secara konsisten dengan *random seed* 42 pada seluruh model untuk menjamin reproduksibilitas eksperimen. Ketidakseimbangan kelas ditangani melalui class_weight="balanced" (SVM, RF, BiLSTM), scale_pos_weight (XGBoost), dan weighted cross-entropy loss dengan bobot dari compute_class_weight scikit-learn (IndoBERT, XLM-R). Model Transformer dioptimasi dengan AdamW dan early stopping. Eksperimen dijalankan satu kali (single-run) menggunakan scikit-learn 1.6.1, XGBoost 3.2.0, TensorFlow 2.20.0, PyTorch 2.11.0 pada GPU Tesla T4, dengan embedding FastText (cc.id.300.vec, Facebook AI Research).

Evaluasi

Hasil keenam model dibandingkan pada tahap perbandingan model untuk menentukan pendekatan terbaik. Evaluasi menggunakan precision, recall, F1-Score, AUC-ROC, dan PR-AUC (Average Precision) sebagai metrik utama pada data tidak seimbang (Lu et al., 2022); *AUC-ROC* mengukur diskriminasi model tanpa bergantung threshold, sedangkan PR-AUC dilaporkan berdampingan karena lebih tepat pada proporsi kelas positif yang kecil (Jamshidian, 2023; Lu et al., 2022). Brier Score dan Expected Calibration Error (ECE) digunakan sebagai metrik kalibrasi sekunder, sementara bootstrap 95% confidence interval dan uji signifikansi berpasangan antar model menilai reliabilitas dan signifikansi statistik selisih performa. Confusion matrix digunakan untuk menganalisis distribusi hasil klasifikasi dan jenis kesalahan prediksi setiap model.

HASIL DAN PEMBAHASAN

Pengujian dilakukan terhadap enam model yang mewakili tiga pendekatan berbeda dalam klasifikasi teks, yaitu *machine learning* tradisional berbasis TF-IDF (SVM, Random Forest, XGBoost), *deep learning* berbasis sekuens (BiLSTM), dan *pretrained language model* berbasis *transformer* (IndoBERT dan XLM-R). Seluruh model dievaluasi pada data uji yang sama menggunakan *threshold* optimal yang diperoleh dari *validation set*. Ringkasan hasil evaluasi disajikan pada Tabel 3.

Tabel 3. Perbandingan Performa Enam Model Klasifikasi pada Data Uji (*Threshold* Optimal)

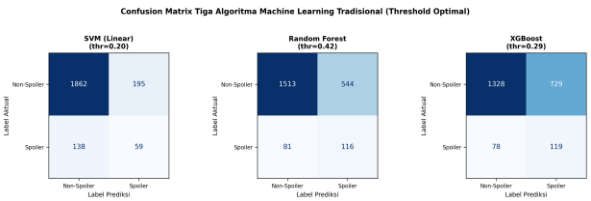
Model	AUC-ROC	PR-AUC	F1 *Spoiler*	*Precision Spoiler*	*Recall Spoiler*
XLM-R	0,7017	0,2263	0,2903	0,2428	0,3609
Indo BERT	0,6939	0,1970	0,2356	0,2328	0,2385
SVM	0,6761	0,1686	0,2214	0,1574	0,3731
RF	0,6675	0,1681	0,2138	0,1786	0,2661
XG Boost	0,6426	0,1605	0,2063	0,1205	0,7156
Bi LSTM	0,6274	0,1658	0,2392	0,1939	0,3119

Sumber: (Hasil Penelitian, 2026)

Berdasarkan Tabel 3, XLM-R memperoleh nilai tertinggi pada AUC-ROC, PR-AUC, dan F1-score *spoiler*, diikuti oleh IndoBERT. Sementara itu, SVM, Random Forest, XGBoost, dan BiLSTM menghasilkan nilai yang lebih rendah pada sebagian besar metrik evaluasi. Selain metrik evaluasi utama, setiap model juga menunjukkan kombinasi *precision* dan *recall* yang berbeda sehingga interpretasi performa tidak dapat didasarkan pada satu metrik saja. Oleh karena itu, pembahasan selanjutnya menganalisis karakteristik masing-masing model melalui kurva ROC, kurva *Precision-Recall*, *threshold*, kalibrasi probabilitas, dan *bootstrap confidence interval*.

Performa Algoritma *Machine Learning* Tradisional

Di antara ketiga algoritma *machine learning* tradisional, SVM memperoleh nilai AUC-ROC dan PR-AUC tertinggi, Random Forest memperoleh F1-score *spoiler* tertinggi, sedangkan XGBoost memperoleh *recall spoiler* tertinggi. Hasil ini menunjukkan bahwa tidak ada satu algoritma yang unggul pada seluruh metrik evaluasi.



Sumber: (Hasil Penelitian, 2026)

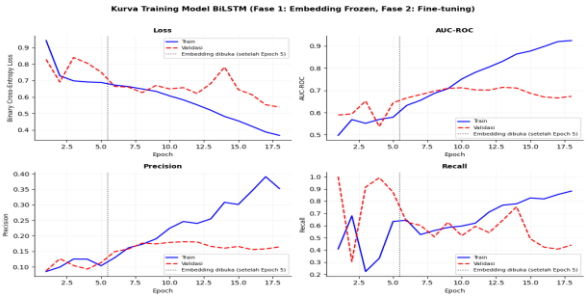
Gambar 2. *Confusion Matrix Algoritma Machine Learning Tradisional (Threshold Optimal)* pada Data Uji (Sumbu X: Label Prediksi; Sumbu Y: Label Aktual)

INTI NUSA MANDIRI
DOI: https://doi.org/10.33480/inti.v21i1.8587
VOL. 21. NO. 1 AGUSTUS 2026
P-ISSN: 0216-6933 | E-ISSN: 2685-807X
Diterbitkan Oleh: LPPM Universitas Nusa Mandiri

Gambar 2 menunjukkan bahwa ketiga algoritma memiliki karakteristik prediksi yang berbeda. SVM cenderung menghasilkan false positive yang lebih sedikit, XGBoost mendeteksi lebih banyak ulasan spoiler tetapi juga menghasilkan false positive yang lebih tinggi, sedangkan Random Forest memberikan keseimbangan yang lebih baik antara precision dan recall. Oleh karena itu, pemilihan model perlu mempertimbangkan lebih dari satu metrik evaluasi.

Performa Model Deep Learning

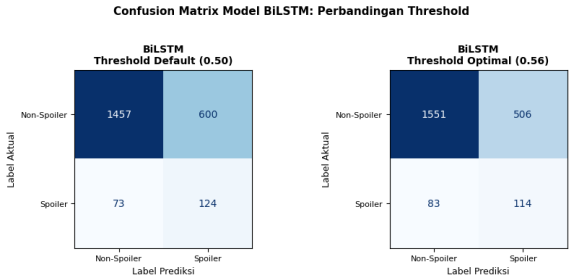
BiLSTM merupakan model deep learning berbasis sekuens yang memanfaatkan pre-trained FastText embeddings. Berdasarkan Tabel 3, performanya masih berada di bawah kedua model transformer, tetapi relatif sebanding dengan algoritma machine learning tradisional pada beberapa metrik evaluasi.



Sumber: (Hasil Penelitian, 2026)

Gambar 3. Kurva Training (Sumbu X: Epoch, Sumbu Y: Metrik Evaluasi) Model BiLSTM pada Data Latih dan Validasi

Gambar 3 menunjukkan bahwa pelatihan dilakukan dalam dua fase, yaitu embedding frozen selama lima Epoch yang dilanjutkan dengan fine-tuning. Selama proses pelatihan, *training loss* terus menurun, sedangkan metrik pada data validasi berfluktuasi setelah beberapa Epoch. Model yang digunakan untuk evaluasi merupakan model dengan nilai AUC terbaik pada validation set.

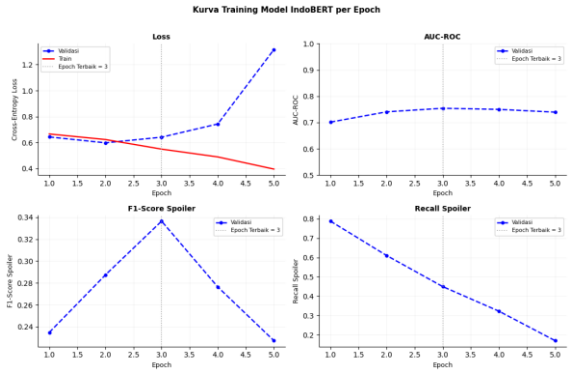


Sumber: (Hasil Penelitian, 2026)

Gambar 4. *Confusion Matrix Model BiLSTM pada Data Uji dengan Threshold Default dan Optimal* (Sumbu X: Label Prediksi; Sumbu Y: Label Aktual)

Gambar 4 menunjukkan bahwa penggunaan *threshold* optimal meningkatkan deteksi ulasan spoiler, tetapi juga meningkatkan jumlah false positive dibandingkan *threshold default*. Hal ini menunjukkan adanya *trade-off* antara kemampuan mendeteksi spoiler dan kesalahan prediksi positif.

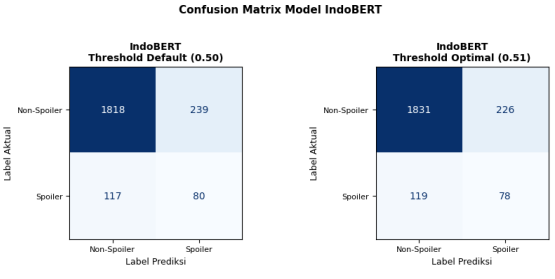
Berbeda dengan BiLSTM, IndoBERT merupakan model transformer yang telah melalui proses pre-training pada korpus bahasa Indonesia. Berdasarkan Tabel 3, IndoBERT memperoleh performa yang lebih tinggi dibandingkan BiLSTM pada sebagian besar metrik evaluasi.



Sumber: (Hasil Penelitian, 2026)

Gambar 5. Kurva Training (Sumbu X: *Epoch*, Sumbu Y: Metrik Evaluasi) Model IndoBERT per *Epoch* pada Data Latih dan Validasi

Gambar 5 menunjukkan bahwa *training loss* terus menurun selama pelatihan, sedangkan metrik pada data validasi mencapai nilai terbaik pada awal pelatihan kemudian cenderung stabil atau berfluktuasi. Model yang digunakan untuk evaluasi merupakan model dengan nilai AUC terbaik pada validation set, yaitu pada Epoch keempat.



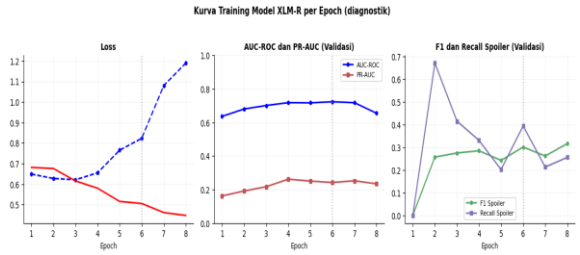
Sumber: (Hasil Penelitian, 2026)

Gambar 6. *Confusion Matrix* Model IndoBERT pada Data Uji dengan *Threshold Default* dan Optimal (Sumbu X: Label Prediksi; Sumbu Y: Label Aktual)

Gambar 6 menunjukkan bahwa penggunaan *threshold* optimal meningkatkan deteksi ulasan spoiler, tetapi juga meningkatkan jumlah false positive dibandingkan *threshold*

VOL. 21. NO. 1 AGUSTUS 2026
P-ISSN: 0216-6933 | E-ISSN: 2685-807X
Diterbitkan Oleh: LPPM Universitas Nusa Mandiri
INTI NUSA MANDIRI
DOI: https://doi.org/10.33480/inti.v21i1.8587

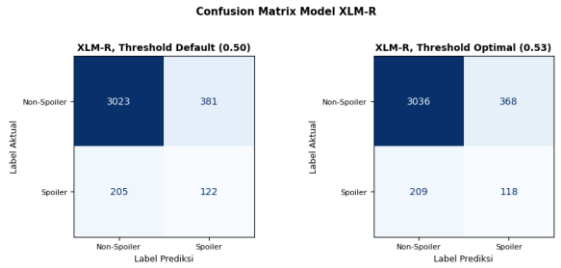
default. Hal ini menunjukkan adanya *trade-off* antara kemampuan mendeteksi spoiler dan kesalahan prediksi positif. XLM-R merupakan model transformer multibahasa yang telah melalui proses pre-training pada berbagai bahasa. Berdasarkan Tabel 3, XLM-R memperoleh nilai tertinggi pada metrik evaluasi utama dibandingkan model lainnya.



Sumber: (Hasil Penelitian, 2026)

Gambar 7. Kurva Training (Sumbu X: *Epoch*, Sumbu Y: Nilai *Loss* dan Metrik Evaluasi) Model XLM-R per *Epoch*

Gambar 7 menunjukkan bahwa training loss terus menurun selama pelatihan, sedangkan metrik pada data validasi mencapai nilai terbaik pada awal pelatihan kemudian berfluktuasi. Model yang digunakan untuk evaluasi merupakan model dengan nilai AUC terbaik pada validation set, yaitu pada Epoch keenam.



Sumber: (Hasil Penelitian, 2026)

Gambar 8. *Confusion Matrix* Model XLM-R pada Data Uji dengan *Threshold Default* dan Optimal

Gambar 8 menunjukkan bahwa penggunaan threshold optimal menghasilkan perubahan hasil klasifikasi yang relatif kecil dibandingkan threshold default. Hal ini menunjukkan bahwa penyesuaian threshold pada XLM-R tidak mengubah keseimbangan antara deteksi spoiler dan kesalahan prediksi secara signifikan.

Analisis *Threshold* dan Kalibrasi

Selain metrik klasifikasi, penelitian ini juga mengevaluasi *threshold* optimal serta kualitas kalibrasi probabilitas setiap model. Ringkasan hasil evaluasi disajikan pada Tabel 4.

Tabel 4. Hasil Evaluasi *Threshold* dan Kalibrasi Probabilitas

Model	Threshold	ECE	Brier
SVM	0,0643	0,0558	0,0829
Random Forest	0,4610	0,2894	0,1619
XGBoost	0,2345	0,1700	0,1083
BiLSTM	0,4200	0,1471	0,1213
IndoBERT	0,3500	0,0554	0,0911
XLM-R	0,5300	0,1057	0,1285

Sumber: (Hasil Penelitian, 2026)

Berdasarkan Tabel 4, setiap model menghasilkan *threshold* optimal yang berbeda. Hal ini menunjukkan bahwa batas keputusan terbaik bergantung pada karakteristik probabilitas yang dihasilkan oleh masing-masing model sehingga penggunaan *threshold* default sebesar 0,50 tidak selalu memberikan performa terbaik. Kualitas kalibrasi probabilitas dievaluasi menggunakan *Expected Calibration Error* (ECE) dan Brier Score, di mana nilai yang lebih rendah menunjukkan probabilitas prediksi yang lebih terkalibrasi. Berdasarkan hasil evaluasi, SVM dan IndoBERT memperoleh nilai ECE terendah, sedangkan SVM memperoleh Brier Score terendah. Hasil ini menunjukkan bahwa performa klasifikasi yang tinggi tidak selalu diikuti oleh kualitas kalibrasi probabilitas yang terbaik.

Analisis Signifikansi Statistik

Selain membandingkan nilai metrik evaluasi, penelitian ini juga menggunakan *bootstrap confidence interval* untuk mengestimasi ketidakpastian nilai AUC-ROC setiap model. Ringkasan hasil estimasi disajikan pada Tabel 5.

Tabel 5. Hasil *Bootstrap Confidence Interval* AUC-ROC Setiap Model

Model	AUC-ROC	95% *Confidence Interval*
SVM	0,6761	[0,6462, 0,7056]
Random Forest	0,6675	[0,6380, 0,6974]
XGBoost	0,6426	[0,6100, 0,6747]
BiLSTM	0,6274	[0,5938, 0,6590]
IndoBERT	0,6939	[0,6615, 0,7233]
XLM-R	0,7017	[0,6700, 0,7329]

Sumber: (Hasil Penelitian, 2026)

Berdasarkan Tabel 5, seluruh model menghasilkan *confidence interval* yang relatif sempit. Namun, sebagian besar *confidence interval* antar model masih saling tumpang tindih sehingga perbedaan nilai AUC-ROC perlu diinterpretasikan secara hati-hati. Untuk menguji signifikansi statistik dari selisih performa antar

model secara eksplisit, dilakukan *paired bootstrap test* (2.000 *resample*) pada data uji untuk ketiga model *machine learning* tradisional. Hasil pengujian menunjukkan bahwa selisih AUC-ROC antara SVM dan Random Forest tidak signifikan secara statistik (selisih = +0,0086; 95% CI [-0,0073, +0,0249]; p = 0,285). Sebaliknya, selisih AUC-ROC antara SVM dan XGBoost (selisih = +0,0335; 95% CI [+0,0097, +0,0592]; p = 0,006) serta antara Random Forest dan XGBoost (selisih = +0,0249; 95% CI [+0,0039, +0,0469]; p = 0,017) terbukti signifikan secara statistik (p < 0,05). Hasil ini mengindikasikan bahwa meskipun XGBoost memperoleh *recall spoiler* tertinggi, performa diskriminatif keseluruhannya (AUC-ROC) secara statistik lebih rendah dibandingkan SVM dan Random Forest. Meskipun demikian, XLM-R tetap memperoleh nilai AUC-ROC tertinggi secara deskriptif berdasarkan hasil evaluasi pada data uji.

Analisis Komparatif, Implikasi, dan Prospek Pengembangan

Pertanyaan berikutnya adalah seberapa kompetitif hasil ini jika diletakkan dalam konteks penelitian deteksi spoiler yang sudah ada. Studi-studi yang relevan semuanya menggunakan dataset Goodreads berbahasa Inggris dengan unit analisis tingkat kalimat dan proporsi spoiler hanya sekitar 3%, melaporkan AUC-ROC antara 0,88 hingga 0,938 (Bao et al., 2021; Chang et al., 2021; Wróblewska et al., 2021). Meskipun lebih rendah, hasil penelitian ini diperoleh pada kondisi eksperimen yang jauh lebih menantang, realistis, dan kokoh secara metodologis.

Pertama, berbeda dengan mayoritas penelitian terdahulu yang membagi ulasan secara acak pada tingkat ulasan atau kalimat sehingga model cenderung mengalami *overfitting* dengan menghafal entitas buku, penelitian ini menerapkan penguncian berbasis kelompok (*group-preserved*) menggunakan variabel *book_url*. Konsekuensinya, model dievaluasi menggunakan ulasan dari buku-buku baru yang sama sekali tidak pernah ditemui selama fase pelatihan. Metode ini menjamin bahwa performa yang dilaporkan mencerminkan kemampuan generalisasi model yang sebenarnya terhadap buku baru. Kedua, penelitian ini menggunakan teks ulasan penuh (bukan kalimat pendek yang terisolasi), menghadapi ketidakseimbangan kelas yang ekstrem (1:10,4), serta menggunakan korpus bahasa lokal (bahasa Indonesia).

Selain itu, F1-score spoiler yang diperoleh seluruh model menunjukkan bahwa deteksi spoiler masih merupakan tugas yang menantang karena spoiler bergantung pada konteks narasi, bukan hanya kata-kata tertentu (Wróblewska et al., 2021). Oleh karena itu, hasil penelitian ini dapat menjadi rujukan awal bagi penelitian deteksi spoiler berbahasa Indonesia selanjutnya.

Keterbatasan Penelitian

Penelitian ini memiliki empat keterbatasan utama. Pertama, meskipun telah dilakukan validasi manual terhadap 300 sampel oleh dua annotator independen dengan tingkat kesepakatan antar-annotator yang sangat tinggi (Cohen's Kappa = 0,963), kesepakatan antara label bawaan Goodreads dan hasil penilaian manusia hanya berada pada level *fair* (Cohen's Kappa = 0,367 dan 0,340 terhadap masing-masing annotator; tingkat kesepakatan mentah 68,3% dan 67,0%). Hal ini mengindikasikan bahwa sebagian label spoiler bawaan Goodreads berpotensi tidak sepenuhnya merepresentasikan konsensus penilaian manusia, sehingga label yang digunakan dalam pelatihan dan evaluasi model tetap mengandung ketidakpastian yang perlu dipertimbangkan dalam interpretasi hasil, dan berpotensi turut memengaruhi batas atas performa yang dapat dicapai oleh seluruh model yang diuji. Kedua, ketidakseimbangan kelas pada dataset (spoiler jauh lebih sedikit dibandingkan non-spoiler) berpotensi memengaruhi kemampuan model mempelajari kelas minoritas.

Ketiga, penggunaan satu dataset dan domain (ulasan buku Goodreads) membatasi generalisasi model pada domain lain. Keempat, prosedur seleksi hyperparameter tidak sepenuhnya setara antara model *machine learning* tradisional dan model *deep learning*/Transformer, sehingga perbandingan performa antar kelompok model perlu diinterpretasikan dengan mempertimbangkan perbedaan prosedur tuning tersebut. Penelitian selanjutnya dapat mengeksplorasi teknik penanganan ketidakseimbangan kelas (data augmentation, few-shot learning), pengujian pada dataset/domain lain, serta perbandingan dengan model terbaru atau Large Language Models (LLMs).

KESIMPULAN

Enam model klasifikasi teks, yaitu SVM, Random Forest, XGBoost, BiLSTM, IndoBERT, dan XLM-R, dibandingkan untuk mendeteksi spoiler pada ulasan buku berbahasa Indonesia dari Goodreads dengan kondisi ketidakseimbangan kelas. Berdasarkan hasil evaluasi, XLM-R memperoleh performa terbaik dengan AUC-ROC sebesar 0,7017, PR-AUC sebesar 0,2263, dan F1-score spoiler sebesar 0,2903, diikuti oleh IndoBERT dengan AUC-ROC sebesar 0,6939. Sementara itu, SVM memperoleh performa terbaik di antara algoritma machine learning tradisional,

menunjukkan bahwa pendekatan berbasis TF-IDF masih mampu memberikan hasil yang kompetitif. Selain itu, hasil penelitian menunjukkan bahwa setiap model memiliki karakteristik prediksi yang berbeda sehingga evaluasi tidak dapat didasarkan pada satu metrik saja. Perbedaan *threshold* optimal antar model juga menunjukkan bahwa pemilihan *threshold* memengaruhi keseimbangan antara *precision* dan *recall*, sedangkan hasil analisis *bootstrap* menunjukkan bahwa sebagian besar perbedaan performa antar model masih memiliki *confidence interval* yang saling tumpang tindih. Oleh karena itu, interpretasi hasil sebaiknya mempertimbangkan metrik evaluasi, *threshold*, dan ketidakpastian statistik secara bersamaan.

REFERENSI

Al-Habib, H., Imah, E. M., Puspitasari, R. D. I., & Prahani, B. K. (2023). *Text Processing Using Support Vector Machine for Scientific Research Paper Content Classification* (pp. 273–282). https://doi.org/10.2991/978-94-6463-174-6_20

Bao, A., Ho, M., & Sangamnerkar, S. (2021). *Spoiler Alert: Using Natural Language Processing to Detect Spoilers in Book Reviews*. http://arxiv.org/abs/2102.03882

Brown, D. W. R. (2025). Spoilers, Narrative Pleasure, and Television. *Projections (New York)*, *19*(2), 25–43. https://doi.org/10.3167/proj.2025.190203

Chang, B., Lee, I., Kim, H., & Kang, J. (2021). *"Killing Me" Is Not a Spoiler: Spoiler Detection Model using Graph Neural Networks with Dependency Relation-Aware Attention Mechanism*. http://arxiv.org/abs/2101.05972

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). *Unsupervised Cross-lingual Representation Learning at Scale*. https://github.com/facebookresearch/cc

Jamshidian, M. (2023). *Evaluation of Text Transformers for Classifying Sentiment of Evaluation of Text Transformers for Classifying Sentiment of Reviews by Using TF-IDF, BERT (word embedding), SBERT Reviews by Using TF-IDF, BERT (word embedding), SBERT (sentence embedding) with Support Vector Machine Evaluation (sentence embedding) with Support Vector Machine Evaluation*. https://arrow.tudublin.ie/scschcomdis

Karakaya, O., & Kilimci, Z. H. (2024). An efficient consolidation of word embedding and deep learning techniques for classifying anticancer peptides: FastText+BiLSTM. *PeerJ Computer Science*, *10*. https://doi.org/10.7717/peerj-cs.1831

Koto, F., Rahimi, A., Lau, J. H., & Baldwin, T. (2020). *IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP*. http://arxiv.org/abs/2011.00677

Lestari, V. B. , U. E. , & H. (2024). Combining Bi-LSTM And Word2vec Embedding For Sentiment Analysis Models of Application User Reviews. *Indonesian Journal of Computer Science*.

Lu, H., Ehwerhemuepha, L., & Rakovski, C. (2022). A comparative study on deep learning models for text classification of unstructured medical notes with various levels of class imbalance. *BMC Medical Research Methodology*, *22*(1). https://doi.org/10.1186/s12874-022-01665-y

Lubis, A. R., Lase, Y. Y., Rahman, D. A., & Witarsyah, D. (2023). Improving Spell Checker Performance for Bahasa Indonesia Using Text Preprocessing Techniques with Deep Learning Models. *Ingenierie Des Systemes d'Information*, *28*(5), 1335–1342. https://doi.org/10.18280/isi.280522

Suhaeni, C., Kamila, S. A., Fahira, F., Yusran, M., & Alfa Dito, G. (2025). Exploring a Large Language Model on the ChatGPT Platform for Indonesian Text Preprocessing Tasks. *Indonesian Journal of Statistics and Its Applications*, *9*(1), 100–116. https://doi.org/10.29244/ijsa.v9i1p100-116

Wróblewska, A., Rzepiński, P., & Sysko-Romańczuk, S. (2021). *Spoiler in a Textstack: How Much Can Transformers Help?* http://arxiv.org/abs/2112.12913

Wyawhare, A. (2023). *Comparative Analysis of Multilingual Text Classification & Identification through Deep Learning and Embedding Visualization*.

Zevana, A., & Riana, D. (2024). TEXT CLASSIFICATION USING INDOBERT FINE-TUNING MODELING WITH CONVOLUTIONAL NEURAL NETWORK AND BI-LSTM. *Jurnal Teknik Informatika (Jutif)*, *4*(6), 1605–1610. https://doi.org/10.52436/1.jutif.2023.4.6.1650

Zhang, L. (2025). Features extraction based on Naive Bayes algorithm and TF-IDF for news classification. *PLOS ONE*, *20*(7 July). https://doi.org/10.1371/journal.pone.0327347

