Sistemasi: Jurnal Sistem Informasi
Volume 15, Nomor 11, 2026: 4787-4801
ISSN:2302-8149
e-ISSN:2540-9719

Klasifikasi *Genre* Novel pada Data tidak Seimbang menggunakan *Machine Learning* dan *Transformer*

Novel Genre Classification on Imbalanced Data using Machine Learning and Transformers

[1]Dania Qomariyah, [2]Muljono*
[1,2]Program Studi Teknik Informatika, Fakultas Teknologi Cerdas, Universitas Dian Nuswantoro
[1,2]Jl. Imam Bonjol No.207, Pendrikan Kidul, Kota Semarang, Jawa Tengah, Indonesia
*e-mail: muljono@dsn.dinus.ac.id

(***received*:** 24 October 2026, ***revised*:** 16 November 2026, ***accepted*:** 17 November 2026)

Abstrak

Klasifikasi genre novel dari sinopsis mendukung pencarian dan rekomendasi bacaan, tetapi distribusi genre pada data nyata umumnya tidak seimbang. Penelitian ini membandingkan Random Forest, Regresi Logistik, SVM linear, dan BERT pada 4.446 baris bersih dari 4.657 baris setelah penghapusan label konflik dan duplikasi. Model klasik memakai TF-IDF 500 fitur serta TF-IDF tanpa batas fitur dengan *bigram*, sedangkan BERT diuji pada 256 dan 512 token dengan *cross-entropy* berbobot dan tanpa bobot kelas. Evaluasi memakai pembagian tetap 70/15/15 dan validasi silang *stratified* k=3, 5, dan 10. Pada pembagian tetap, *random resampling* menurunkan akurasi Random Forest dari 60,42% menjadi 45,88% tanpa meningkatkan F1-macro, sedangkan pembobotan kelas menaikkan F1-macro dari 0,424 menjadi 0,586. Pada validasi silang, BERT 512 token berbobot kelas memperoleh rata-rata F1-macro tertinggi (0,809-0,816) dan akurasi 77,3-78,7%, dibandingkan SVM dengan 0,752-0,766 dan 75,1-76,1%. Pada k=10, selisih F1 terbesar terdapat pada romance, travel, dan sports, tetapi peningkatan juga tampak pada genre yang lebih besar. Pada ketiga nilai k, waktu latih per *fold* BERT tersebut sekitar 5.300-6.600 kali waktu SVM (hanya waktu pelatihan). SVM dengan TF-IDF tanpa batas fitur tetap menjadi pembanding kuat yang efisien. Hasil ini menunjukkan peningkatan kinerja BERT disertai tambahan waktu pelatihan yang besar, sehingga pemilihan model perlu mempertimbangkan prioritas aplikasi.
Kata kunci: ketidakseimbangan kelas, klasifikasi genre novel, random forest, sinopsis, transformer

Abstract

Novel genre classification based on plot summaries can support book search and reading recommendation systems. However, genre distributions in real-world datasets are generally imbalanced. This study compares Random Forest, Logistic Regression, linear Support Vector Machine (SVM), and Bidirectional Encoder Representations from Transformers (BERT) using 4,446 cleaned records out of an initial 4,657 records after removing conflicting labels and duplicate entries. The classical machine learning models employed TF-IDF representations with 500 features and unrestricted features incorporating bigrams, while BERT was evaluated using sequence lengths of 256 and 512 tokens, with both weighted and unweighted cross-entropy loss. Evaluation was conducted using a fixed 70/15/15 train-validation-test split and stratified k-fold cross-validation with k = 3, 5, and 10. Under the fixed split, random resampling reduced Random Forest accuracy from 60.42% to 45.88% without improving macro-F1, whereas class weighting increased macro-F1 from 0.424 to 0.586. In cross-validation, BERT with a 512-token sequence length and class-weighted loss achieved the highest average macro-F1 scores (0.809–0.816) and accuracy (77.3%–78.7%), outperforming SVM, which achieved macro-F1 scores of 0.752–0.766 and accuracy of 75.1%–76.1%. At k = 10, the largest F1-score improvements were observed in the romance, travel, and sports genres, although improvements were also evident in larger genres. Across all three k values, the training time per fold for this BERT configuration was approximately 5,300–6,600 times that of SVM, considering training time only. SVM with unrestricted TF-IDF features remained a strong and computationally efficient baseline. These findings indicate that BERT delivers improved classification

 ISSN:2302-8149
e-ISSN:2540-9719

performance at the cost of substantially higher training time, suggesting that model selection should consider the priorities and computational constraints of the intended application.
Keywords: *class imbalance, novel genre classification, random forest, synopsis, transformer*

1 Pendahuluan

Jumlah buku yang terbit terus bertambah dan peredarannya beralih ke format e-book dan audiobook, sehingga klasifikasi genre otomatis dibutuhkan untuk membantu pembaca menemukan bacaan dan meningkatkan efisiensi sistem rekomendasi [1]. Klasifikasi genre juga berperan dalam pengorganisasian perpustakaan, temu kembali informasi, dan rekomendasi personal [2]. Pelabelan genre secara manual memerlukan waktu dan tenaga yang besar, sehingga dikembangkan klasifikasi otomatis berbasis deskripsi buku [3]. Sinopsis cocok dipakai karena memuat informasi inti tentang tema, alur, dan karakteristik cerita [4].

Pendekatan klasifikasi genre buku berkisar dari model klasik berbasis *term frequency-inverse document frequency* (TF-IDF), seperti Naïve Bayes, Gradient Boosting, dan Random Forest [3], [4], serta Support Vector Machine (SVM) dan Regresi Logistik [1], hingga model berbasis Bidirectional Encoder Representations from Transformers (BERT) [2], [5]. Pada klasifikasi teks secara umum, BERT dan model transformer dilaporkan lebih baik daripada model klasik [6], tetapi hasilnya tidak seragam: pada teks pendidikan, BERT dan RoBERTa justru di bawah Naïve Bayes dan hanya GPT-3 yang melampaui model klasik [7]. Selain itu, pengklasifikasi linear pada fitur *bag-of-words* sering kompetitif dengan efisiensi tinggi sehingga dianjurkan selalu dijalankan sebagai pembanding [8]. Keunggulan BERT juga dapat disertai biaya komputasi besar: pada empat tugas klasifikasi tweet, model berbasis BERT paling akurat, tetapi hanya unggul 0,61% sampai 9,79% atas model non-BERT terbaik, dengan waktu implementasi varian BERT RoBERTa 1.116-1.193 detik dan CamemBERT 1.458-1.493 detik, dibanding 1,74-12,9 detik untuk Random Forest [9].

Tantangan lain adalah distribusi kelas yang tidak merata. Banyak tugas pemrosesan bahasa alami (natural language processing, NLP) tidak seimbang secara alami, dan model cenderung berkinerja buruk pada kelas yang jarang muncul [10]. Dataset yang dipakai memuat sepuluh genre dengan sebaran timpang, misalnya pada data mentah thriller 1.023 sinopsis sedangkan romance hanya 111 [11]. Pada data tidak seimbang, akurasi dapat menyesatkan, sedangkan F1-macro memberi bobot sama pada setiap kelas sehingga kinerja buruk pada kelas minoritas ikut menurunkan skor [12]. Pada sinopsis novel, Random Forest lebih mudah mengenali genre yang datanya banyak, sedangkan kesalahan terutama terjadi pada psychology, sports, dan romance [4]. Penanganannya antara lain *random resampling* 100 sampel per genre [4] dan *focal loss* pada model BERT berbahasa Indonesia (IndoBERT) untuk berita hoaks, yang memberi akurasi 98,29% dibanding 97,97% (selisih 0,32 poin) pada *cross-entropy* [13].

Mahanani dan Fibriani [4] mengevaluasi Random Forest pada dataset Kaggle yang sama (4.657 baris). Mereka memakai subset 1.000 baris, 500 kata berbobot TF-IDF tertinggi, dan pembagian data 80/20. Akurasi naik dari 42,11% menjadi 46,67% setelah *random resampling* 100 sampel per genre pada data latih. Matriks kebingungan (*confusion matrix*) yang dilaporkan memuat sekitar 195 data uji, dan genre psychology, romance, sports, serta travel masing-masing hanya diwakili empat sampel uji. Studi tersebut tidak melaporkan validasi silang, audit kualitas data, maupun pembanding selain Random Forest. Penulisnya menyarankan Word2Vec, FastText, atau BERT sebagai penelitian lanjutan. Adhikari [14] menguji Multinomial Naïve Bayes, Regresi Logistik, SVC, dan Random Forest pada dataset Kaggle yang sama dengan jumlah baris dan genre yang sama, memakai judul dan sinopsis serta satu pembagian 80/20, dan melaporkan akurasi 65,12-67,81% tanpa penanganan ketidakseimbangan kelas, validasi silang, maupun model transformer.

Di sisi lain, sebuah buku dapat termasuk lebih dari satu genre [1], dan Yang dan Zhang mengakui bahwa penetapan satu genre per buku belum sepenuhnya mencerminkan kompleksitas karya sastra modern [2], sehingga konsistensi label pada data genre perlu diperiksa. Berdasarkan uraian tersebut, terdapat beberapa celah. Pertama, kedua studi pada dataset ini [4], [14] memakai satu pembagian data tanpa validasi silang, dan studi [4] hanya memiliki sekitar 195 data uji, sehingga belum jelas apakah peningkatan akibat penyeimbangan konsisten pada pembagian data yang berbeda. Kedua, sepengetahuan penulis, belum ada perbandingan BERT dengan pembanding klasik yang dikonfigurasi kuat (misalnya TF-IDF tanpa batas fitur dengan *bigram*) pada dataset ini, padahal pembanding linear

 ISSN:2302-8149
e-ISSN:2540-9719

dianjurkan [8]. Ketiga, kedua studi tidak melaporkan audit duplikasi dan label konflik [4], [14], padahal duplikasi pada korpus pemodelan bahasa dilaporkan memengaruhi evaluasi [15].

Penelitian ini bertujuan membandingkan model klasik (Random Forest, Regresi Logistik, dan SVM) dengan BERT untuk klasifikasi genre novel dari sinopsis pada data tidak seimbang. Untuk itu, penelitian ini mengaudit dan membersihkan dataset, mengadaptasi prosedur penyeimbangan *random resampling* penelitian terdahulu [4] dan membandingkannya dengan kondisi tanpa penyeimbangan serta pembobotan kelas, lalu mengevaluasi seluruh model dengan validasi silang *stratified* k=3, 5, dan 10 sebagai hasil utama, dilengkapi pembagian data tetap, analisis kinerja per genre, dan ablasi panjang token serta pembobotan kelas pada BERT. Pembanding linear yang kuat dijalankan sesuai anjuran [8]. Pertanyaan penelitian ini: (1) apakah *random resampling* memperbaiki F1-macro dibanding kondisi tanpa penanganan maupun pembobotan kelas; (2) seberapa besar selisih BERT terhadap pembanding klasik terbaik, dan apakah konsisten pada k=3, 5, dan 10; (3) seberapa besar waktu latih BERT dibanding model klasik; dan (4) apakah peningkatan panjang masukan BERT dari 256 menjadi 512 token serta penggunaan pembobotan kelas memengaruhi F1-macro dan kinerja genre minoritas.

2 Tinjauan Literatur

Penelitian klasifikasi genre buku umumnya memakai model klasik dengan TF-IDF pada data berukuran kecil sampai menengah. Sethy dkk. [3] membandingkan Multinomial Naïve Bayes, Random Forest, dan Gradient Boosting pada sekitar 3.000 ringkasan buku dari dataset CMU dengan enam genre berukuran sama (500 baris per genre). Pada pembagian 80:20, Naïve Bayes mencapai akurasi 78%, di atas Random Forest (65,66%) dan Gradient Boosting (65,33%). Karena kelasnya seimbang, hasil tersebut tidak menjawab bagaimana model berperilaku pada distribusi genre yang timpang. Mahanani dan Fibriani [4] memakai dataset Kaggle yang tidak seimbang dan melaporkan kenaikan akurasi Random Forest dari 42,11% menjadi 46,67% setelah *random resampling* 100 sampel per genre, tetapi pada subset 1.000 baris, dengan satu pembagian 80:20, dan hanya satu algoritma. Pada dataset Kaggle yang sama, Adhikari [14] menguji Multinomial Naïve Bayes, Regresi Logistik, SVC, dan Random Forest memakai judul dan sinopsis dengan satu pembagian 80:20 dan tanpa penanganan ketidakseimbangan kelas; Regresi Logistik memberi F1 tertinggi (0,660) meski akurasinya (67,06%) di bawah Random Forest (67,81%; F1 hanya 0,541), menegaskan bahwa akurasi tertinggi tidak selalu berarti F1 tertinggi pada data yang timpang. Subroto dan Faisal [1] memodelkan genre sebagai masalah *multilabel* pada 8.814 buku Goodreads dengan 50 genre teratas dan memakai validasi silang 10-fold. SVM dengan *stemming* terbaik pada semua metrik (akurasi 0,5197; F1 0,7692), dan Label Powerset memberi akurasi lebih tinggi (0,5423) daripada Binary Relevance (0,5197). Penulisnya mengaitkan akurasi di bawah 80% dengan distribusi data yang tidak merata. Angka-angka ini tidak dapat dibandingkan langsung karena dataset, jumlah genre, dan definisi tugasnya berbeda, tetapi keempatnya menunjukkan bahwa kinerja klasifikasi genre sangat dipengaruhi komposisi data dan metrik yang dipakai.

Representasi kontekstual sudah diuji pada tugas genre, tetapi dengan masukan selain sinopsis. Yang dan Zhang [2] memakai 57.000 buku dengan 30 genre. BERT yang hanya memakai judul memperoleh *balanced accuracy* 0,6109 dan F1 0,5866, jauh di bawah model citra sampul InceptionV3 (0,7895 dan 0,7845), sedangkan gabungan keduanya mencapai 0,7951 dan 0,7920. Mereka sendiri menyebut penetapan satu genre per buku sebagai keterbatasan karena banyak buku berada pada beberapa genre. Hamaguchi dkk. [5] menunjukkan bahwa BERT mengungguli *continuous bag of words* (CBOW) terutama ketika bentuk teks mirip: akurasi tertinggi 92,25% pada pasangan novel-esai, sedangkan pada novel-puisi yang bentuknya berbeda jauh kedua model sama-sama mencapai sekitar 99%. Solovyev dkk. [16] melaporkan akurasi 95,9% dari model terbaik (rubert-tiny) pada 743 dongeng Rusia. Temuan-temuan ini menunjukkan bahwa representasi kontekstual layak untuk tugas genre, tetapi masukannya adalah judul, sampul, atau teks sastra, bukan sinopsis novel.

Pada klasifikasi teks umum, bukti tentang keunggulan transformer beragam. Garrido-Merchán dkk. [6] melaporkan BERT lebih baik daripada model berbasis TF-IDF dan AutoML, misalnya pada ulasan IMDB (0,9387 dibanding 0,9007 untuk voting classifier), berita berbahasa Portugis (0,9093 dibanding 0,8480 untuk AutoML), dan ulasan hotel berbahasa Cina (0,9381 dibanding 0,7399 untuk AutoML). Pola serupa dilaporkan pada teks berbahasa Indonesia: Arroyan dkk. [17] mendapati

 ISSN:2302-8149
e-ISSN:2540-9719

IndoBERT 98,70% pada komentar judi online, di atas SVM (88,85%) dan Random Forest (79,62%), dan Medantoro dan Muljono [18] melaporkan IndoBERT 97% pada cuitan kebijakan pendidikan, di atas Random Forest dan SVM (masing-masing 95%). Wang dkk. [19], yang membandingkan sepuluh model pada klasifikasi sentimen lima kelas atas 9.382 artikel berita, juga mendapati BERT berkinerja baik secara keseluruhan. Namun, bukti lain memberi gambaran yang lebih bernuansa. Lin dkk. [8] menunjukkan bahwa pengklasifikasi linear pada fitur *bag-of-words* sering kompetitif dengan efisiensi tinggi, sehingga mereka menganjurkan pembanding linear selalu dijalankan. Aydın dkk. [7], pada teks pendidikan enam kategori, mendapati Generative Pre-trained Transformer 3 (GPT-3) terbaik (akurasi 0,77), tetapi BERT (0,50), DistilBERT (0,50), RoBERTa (0,56), dan GPT-2 (0,44) justru di bawah Naïve Bayes (0,65), Random Forest (0,61), dan SVM (0,59). Siino dkk. [20], pada empat dataset dan sembilan model, menunjukkan bahwa pilihan prapemrosesan dapat mengubah akurasi hingga 25% dan pada beberapa kasus Naïve Bayes sederhana melampaui transformer terbaik sebesar 2% akurasi. Hassan dkk. [21] mendapati peringkat model bergantung pada dataset: Regresi Logistik terbaik pada IMDB (85,8%) dan k-nearest neighbors pada data spam (98,5%). Benítez-Andrades dkk. [9], pada empat tugas biner tweet gangguan makan, mendapati model berbasis BERT terbaik, tetapi dengan peningkatan akurasi hanya 0,61% sampai 9,79% atas model sederhana terbaik, dengan waktu implementasi RoBERTa 1.116-1.193 detik dan CamemBERT 1.458-1.493 detik, dibanding 1,74-12,9 detik untuk Random Forest. Mereka menilai biaya komputasi tersebut mungkin berlebihan. Secara keseluruhan, pola-pola yang kontras ini menegaskan bahwa keunggulan transformer bersifat spesifik-dataset dan spesifik-tugas, sehingga pembanding klasik yang kuat tetap perlu disertakan sebelum menyimpulkan keunggulan BERT pada tugas baru seperti klasifikasi genre novel.

Ketidakseimbangan kelas menambah kerumitan perbandingan tersebut. Henning dkk. [10] mengelompokkan metode penanganannya di NLP menjadi berbasis sampling, augmentasi data, pilihan fungsi loss, staged learning, dan desain model, sedangkan Ghosh dkk. [22] menelaah apakah hubungan antara ketidakseimbangan kelas, kompleksitas konsep, ukuran data, dan kinerja yang dikenal pada pembelajaran klasik juga berlaku pada *deep learning*. Taskiran dkk. [23] membandingkan SMOTE dan 30 variannya pada dua dataset teks (TREC dan Emotions) dengan representasi transformer dan memvalidasi perbedaannya dengan uji Friedman. Pada model transformer, Kunaefi dkk. [13] memakai *focal loss* pada IndoBERT untuk berita hoaks yang tidak seimbang dan melaporkan akurasi 0,9829 dibanding 0,9797 untuk *cross-entropy*, dengan *recall* kelas hoaks 0,9646 dibanding 0,9513. Mereka memilihnya karena tidak memerlukan modifikasi data seperti *oversampling* atau *undersampling*. Widayanti dan Kasih [24] membandingkan IndoBERT *baseline*, IndoBERT berbobot kelas, dan SMOTE+SVM pada klasifikasi sentimen tidak seimbang dan mendapatkan F1- macro masing-masing 0,598, 0,582, dan 0,545. IndoBERT berbobot kelas dinilai lebih seimbang pada kelas minoritas meski skor rata-ratanya sedikit lebih rendah daripada *baseline*. Temuan ini menunjukkan bahwa satu skor rata-rata dapat menyembunyikan perbedaan antar kelas, sejalan dengan Hinojosa Lee dkk. [12] yang menekankan pentingnya memilih varian F1 (micro, macro, atau weighted) pada data *multilabel* yang tidak seimbang. Selain itu, kualitas data memengaruhi kesahihan evaluasi: Lee dkk. [15] menunjukkan bahwa tumpang tindih antara data latih dan uji memengaruhi lebih dari 4% set validasi pada dataset standar. Secara bersama, temuan-temuan ini menegaskan bahwa penanganan ketidakseimbangan kelas dan kualitas data perlu dievaluasi bersamaan, bukan terpisah, karena audit data yang lemah dapat mengaburkan efek sesungguhnya dari suatu teknik penyeimbangan.

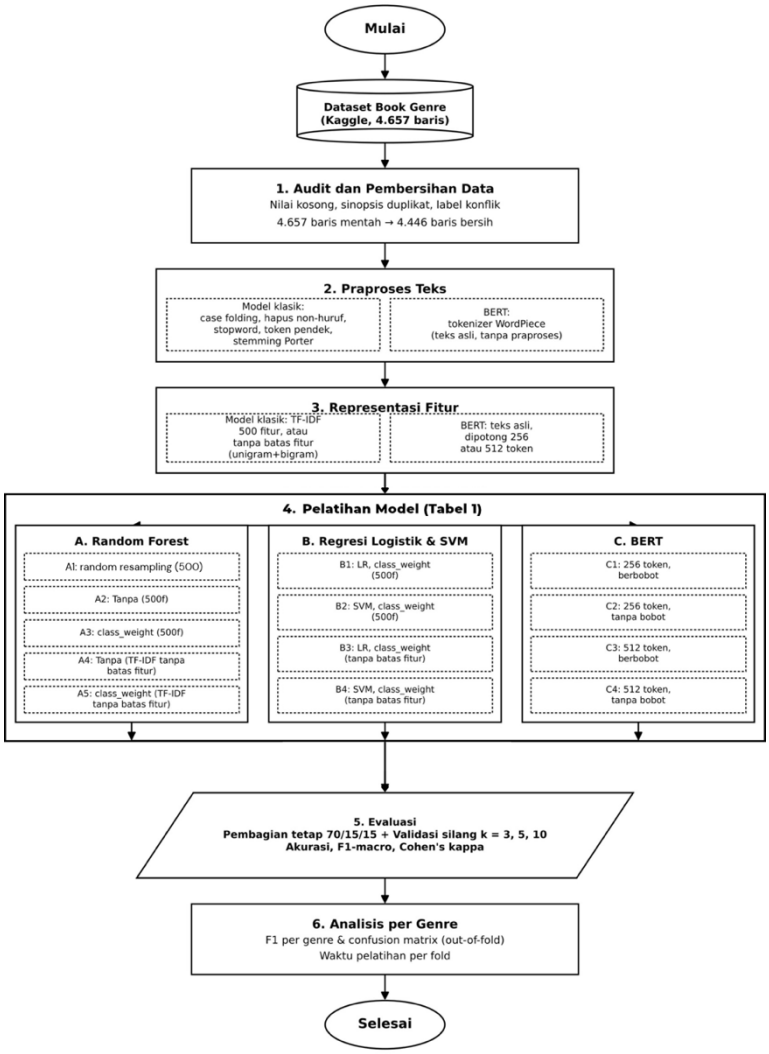
Dari telaah tersebut terdapat beberapa celah yang menjadi fokus penelitian ini. Pertama, kedua studi genre pada dataset Kaggle yang sama [4], [14] memakai satu pembagian data tanpa validasi silang, dengan [4] terbatas pada subset 1.000 baris dan satu algoritma, serta [14] tanpa penanganan ketidakseimbangan kelas sama sekali, sedangkan validasi silang baru muncul pada studi *multilabel* dengan dataset berbeda [1]. Kedua, bukti tentang transformer pada tugas genre berasal dari judul, sampul, atau teks sastra [2], [5], [16], sementara perbandingan model klasik dan transformer pada teks umum bergantung pada dataset dan biaya komputasi [6], [7], [9], [20]. Ketiga, di antara literatur yang ditelaah, tidak ada studi genre yang membandingkan penyeimbangan data, pembobotan kelas, dan kondisi tanpa penanganan pada pembagian data yang sama, atau yang mengaudit duplikasi dan label konflik sebelum evaluasi [15]. Keempat, belum ada studi pada dataset ini yang menguji pengaruh panjang masukan dan pembobotan kelas pada BERT. Penelitian ini mengisi celah tersebut dengan membandingkan Random Forest, Regresi Logistik, dan SVM linear dengan BERT pada seluruh

 ISSN:2302-8149
e-ISSN:2540-9719

dataset sinopsis yang telah diaudit, memakai validasi silang *stratified* k=3, 5, dan 10 dengan fold identik untuk semua model, pembagian data tetap sebagai pelengkap, analisis per genre dengan F1-macro, pencatatan waktu latih, serta ablasi panjang masukan dan pembobotan kelas pada BERT. Pembanding linear yang kuat dijalankan sejalan dengan anjuran Lin dkk. [8].

3 Metode Penelitian

Penelitian ini bersifat kuantitatif eksperimental. Alur penelitian terdiri atas enam tahap: audit dan pembersihan data, prapemrosesan teks, representasi fitur, pelatihan model, evaluasi, dan analisis per genre (Gambar 1).



Gambar 1 Alur penelitian

3.1 Dataset dan Audit Data

Dataset yang dipakai adalah Book Genre Prediction using NLP dari Kaggle [11], berisi 4.657 baris dengan kolom *index*, *title*, *genre*, dan *summary* untuk 10 genre. Berbeda dari penelitian terdahulu [4] yang memakai subset 1.000 baris, penelitian ini memakai seluruh data setelah audit. Audit dilakukan karena tumpang tindih antara data latih dan uji memengaruhi ketepatan evaluasi [15], dan mencakup tiga pemeriksaan: nilai kosong, sinopsis ganda, dan sinopsis yang muncul dengan lebih dari satu label genre. Baris dengan label konflik dihapus terlebih dahulu. Selanjutnya, dari sisa data, sinopsis duplikat berlabel sama disaring sehingga hanya satu representasi dipertahankan per kelompok duplikat. Hasil audit disajikan pada Subbab 4.1, dan seluruh eksperimen memakai data bersih tersebut.

 ISSN:2302-8149
e-ISSN:2540-9719

3.2 Prapemrosesan dan Representasi Data

Untuk model klasik, sinopsis melalui *case folding*, penghapusan karakter non-huruf, tokenisasi berdasarkan spasi, penghapusan *stopword* bahasa Inggris (daftar scikit-learn) dan token berpanjang kurang dari tiga karakter, serta *stemming* Porter. Untuk BERT, sinopsis asli dipakai tanpa prapemrosesan karena model memiliki tokenizer sendiri, dan pengaruh prapemrosesan dilaporkan bergantung pada model [20]. Masukan BERT dibatasi 256 atau 512 token (termasuk [CLS] dan [SEP]) sesuai konfigurasi (Tabel 1): sinopsis yang lebih panjang dipotong, sedangkan yang lebih pendek diberi token pengisi (padding).

Dua representasi TF-IDF dibandingkan. Pertama, 500 fitur (*max_features* = 500), yaitu 500 term berfrekuensi tertinggi pada data latih, mengikuti jumlah fitur penelitian terdahulu [4]. Kedua, TF-IDF tanpa batas max_features, menggunakan unigram dan *bigram*, min_df = 2, dan sublinear_tf = True, yang menghasilkan 57.147 fitur pada data latih pembagian tetap (selanjutnya disebut TF-IDF tanpa batas fitur). Pada kedua representasi, *vectorizer* di-*fit* hanya pada data latih (pada validasi silang, di dalam tiap *fold*) untuk mencegah kebocoran informasi dari data uji.

3.3 Model dan Konfigurasi

Empat algoritma dibandingkan (Tabel 1): Random Forest, Regresi Logistik, SVM linear (LinearSVC), dan BERT (bert-base-uncased) [25]. Random Forest memakai 300 pohon pada seluruh konfigurasi. Model klasik dilatih pada dua representasi TF-IDF, sedangkan BERT memakai teks sinopsis asli. Pada Tabel 5-8, LR = Regresi Logistik, 500 = TF-IDF 500 fitur, dan penuh = TF-IDF tanpa batas fitur.

Tabel 1 Konfigurasi model

Kode	Model	Fitur	Penanganan imbalance	Evaluasi
A1	Random Forest	500 fitur	random resampling [4]	Tetap, CV
A2	Random Forest	500 fitur	tanpa penanganan	Tetap, CV
A3	Random Forest	500 fitur	class_weight	Tetap, CV
A4	Random Forest	TF-IDF tanpa batas fitur	tanpa penanganan	Tetap, CV
A5	Random Forest	TF-IDF tanpa batas fitur	class_weight	Tetap, CV
B1	Regresi Logistik	500 fitur	class_weight	Tetap, CV
B2	SVM linear	500 fitur	class_weight	Tetap, CV
B3	Regresi Logistik	TF-IDF tanpa batas fitur	class_weight	Tetap, CV
B4	SVM linear	TF-IDF tanpa batas fitur	class_weight	Tetap, CV
C1	BERT, 256 token	Teks asli	cross-entropy berbobot	Tetap, CV
C2	BERT, 256 token	Teks asli	cross-entropy tanpa bobot	Tetap, CV
C3	BERT, 512 token	Teks asli	cross-entropy berbobot	Tetap, CV
C4	BERT, 512 token	Teks asli	cross-entropy tanpa bobot	Tetap, CV

Penanganan ketidakseimbangan kelas dibandingkan dalam tiga kondisi: tanpa penanganan, *random resampling* penelitian terdahulu (A1), dan pembobotan kelas. Pada A1, 100 sampel per genre diambil dengan pengembalian dari data latih. Pada pembobotan kelas model klasik (*class_weight* = *balanced*), bobot kelas *c* dihitung dengan Persamaan (1):

$$w_c = \frac{N}{K \times n_c} \tag{1}$$

dengan N jumlah sampel latih, K jumlah kelas, dan n_c jumlah sampel kelas c. Pada BERT, bobot yang sama dipakai dalam *cross-entropy* berbobot pada Persamaan (2):

$$L = -\frac{\sum_{i=1}^{B} w_{y_i} \log p_i(y_i)}{\sum_{i=1}^{B} w_{y_i}} \tag{2}$$

dengan B jumlah sampel dalam satu *batch*, y_i label sebenarnya sampel i, dan $p_i(y_i)$ probabilitas yang diprediksi untuk label tersebut.

Pada BERT, efek pembobotan ini turut diuji sebagai ablasi dengan membandingkan *cross-entropy* berbobot (C1, C3) dan tanpa bobot (C2, C4), sejalan dengan perbandingan tanpa penanganan vs class_weight pada model klasik. BERT di-fine-tune pada dua panjang maksimum token, 256 dan 512, untuk menguji pengaruh panjang konteks terhadap kinerja (C1-C2 vs C3-C4). Hyperparameter: learning rate 2×10^{-5}, ukuran batch 16, weight decay 0,01, warmup 10%, dan maksimum 5 epoch.

 ISSN:2302-8149
e-ISSN:2540-9719

Early stopping dengan kesabaran 2 epoch berdasarkan F1-macro validasi, dan model dengan F1-macro validasi terbaik dipakai untuk prediksi. Seluruh seed diatur 42.

3.4 Pembagian Data Tetap

Pembagian tetap *stratified* 70/15/15 dipakai untuk menjaga proporsi kelas, tetapi ukuran sampel uji yang kecil pada kelas minoritas tetap menjadi perhatian pada evaluasi data tidak seimbang [12]. Data validasi hanya dipakai BERT untuk *early stopping*, sedangkan model klasik dilatih pada data latih saja. Karena kelas minoritas pada pembagian ini hanya memiliki 12 sampai 15 sampel uji (psychology 13, romance 12, sports 12, travel 15), hasil pada skema ini rentan terhadap variasi sehingga dipakai sebagai pelengkap, bukan klaim utama.

3.5 Validasi Silang

Validasi silang dijadikan hasil utama penelitian ini, berbeda dari dua studi genre pada dataset yang sama [4], [14] yang hanya memakai satu pembagian data, dan sejalan dengan pemakaian validasi silang pada studi *multilabel* dataset berbeda [1]. Digunakan validasi silang *stratified* dengan k=3, 5, dan 10, dengan pengacakan dan *seed* tetap (42) sehingga *fold* yang dihasilkan identik untuk semua model pada tiap nilai k. Pada tiap *fold*, BERT membagi data latih *fold* menjadi 85% latih dan 15% validasi (*stratified*) untuk *early stopping*, sedangkan model klasik memakai seluruh data latih fold tanpa pemisahan validasi tambahan. Prediksi *out-of-fold* dari seluruh fold digabungkan untuk menghitung F1 per genre dan *confusion matrix*. Waktu latih per *fold* dicatat untuk membandingkan biaya komputasi antar model. Skema ini dipilih sebagai hasil utama karena ukuran sampel uji yang jauh lebih besar dan stabil dibanding pembagian tetap, terutama pada kelas minoritas.

3.6 Metrik Evaluasi

Kinerja dinilai dengan akurasi, F1-macro, dan Cohen's kappa. F1-macro menjadi metrik utama karena rata-rata tak berbobot antar kelas tidak menutupi kegagalan pada kelas minoritas [12], dengan perhitungan pada Persamaan (3)-(5):

$$F1_c = \frac{2P_cR_c}{P_c+R_c} \tag{3}$$

$$\text{F1} - \text{macro} = \frac{1}{K}\sum_{c=1}^{K} F1_c \tag{4}$$

$$\kappa = \frac{p_o-p_e}{1-p_e} \tag{5}$$

dengan P_c dan R_c presisi dan *recall* kelas c, p_o akurasi teramati, dan p_e akurasi yang diharapkan secara acak. Pada validasi silang, hasil dilaporkan sebagai rata-rata dan simpangan baku sampel antar-*fold*. Rata-rata dan simpangan baku digunakan untuk menggambarkan kinerja serta variasi antar-*fold*. Waktu diukur dengan time.time() sebagai waktu pelatihan: untuk model klasik hanya fit klasifier (TF-IDF, resampling, dan prediksi tidak dihitung), dan untuk BERT trainer.train() beserta validasi tiap epoch (tokenisasi tidak dihitung). Seluruh eksperimen dijalankan pada runtime dengan GPU T4 (di Kaggle, T4 ×2 dibatasi ke satu GPU); BERT memakai GPU, sedangkan model klasik memakai prosesor runtime. Pembagian tetap, k=3, dan k=5 di Google Colab, k=10 di Kaggle. Karena platform dan ukuran data latih per fold berbeda, waktu antar-k tidak dibandingkan; rasio BERT/klasik dihitung pada tiap k di platform yang sama.

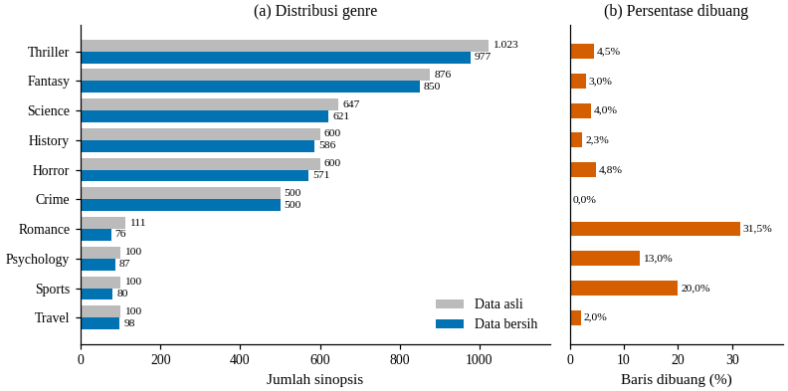
4 Hasil dan Pembahasan

4.1 Hasil Audit Data

Audit menemukan tidak ada nilai kosong, 228 baris (4,9%) terlibat duplikat sinopsis berlabel sama (113 kelompok), dan 96 baris terlibat konflik label, yaitu sinopsis identik yang muncul pada beberapa baris dengan label genre berbeda. Setelah baris berlabel konflik dibuang dan duplikat berlabel sama dikurangi menjadi satu baris, tersisa 4.446 baris bersih, atau 211 baris (4,5%) dibuang. Dari 211 baris yang dibuang, 96 baris berasal dari label konflik dan 115 baris merupakan kelebihan salinan pada kelompok duplikat berlabel sama (228 baris yang terlibat duplikat berasal dari 113 kelompok unik, masing-masing menyisakan satu representasi). Pembuangan tidak merata antar genre (Gambar 2b): genre kecil paling terdampak, yaitu romance 31,5%, sports 20,0%, dan psychology

 ISSN:2302-8149
e-ISSN:2540-9719

13,0%, sedangkan crime tidak terdampak. Duplikasi dan label konflik berpotensi merugikan evaluasi genre minoritas apabila tidak dibersihkan, tetapi dampak keduanya terhadap kinerja model tidak diuji secara terpisah dalam penelitian ini, meski penelitian terdahulu [4] tidak melaporkan audit data serupa untuk pembanding. Setelah pembersihan, rasio genre terbesar (thriller, 977 baris) terhadap terkecil (romance, 76 baris) tetap sekitar 12,9:1 (Gambar 2a), sehingga ketidakseimbangan kelas masih berat. Pemeriksaan akhir memastikan tidak ada sinopsis ganda, data uji yang identik dengan data latih, dan tidak ada irisan indeks antar bagian data.



Gambar 2 Distribusi genre sebelum dan sesudah pembersihan data: (a) jumlah sinopsis per genre, (b) persentase baris yang dibuang

4.2 Hasil Prapemrosesan dan Pembagian Data

Tabel 2 menunjukkan hasil tiap tahap prapemrosesan model klasik pada kalimat pertama salah satu sinopsis. *Case folding* dan penghapusan tanda baca memecah kalimat menjadi 23 token, penghapusan *stopword* dan token pendek menyisakan 9 token, dan *stemming* Porter mengembalikan kata ke bentuk dasarnya, misalnya *opens* menjadi *open* dan *awakens* menjadi *awaken*, meski kadang menghasilkan bentuk tidak utuh, seperti *bus* menjadi *bu* dan *unable* menjadi *unabl*.

Tabel 2 Hasil prapemrosesan model klasik pada satu kalimat sinopsis

Tahap	Hasil
Teks asli	*As the book opens, Jason awakens on a school bus, unable to remember who or where he is, or anything about his past.*
Case folding dan pembersihan	*as the book opens jason awakens on a school bus unable to remember who or where he is or anything about his past*
Tokenisasi (23 token)	*as, the, book, opens, jason, awakens, on, a, school, bus, unable, to, remember, who, or, where, he, is, or, anything, about, his, past*
Penghapusan *stopword* (9 token)	*book, opens, jason, awakens, school, bus, unable, remember, past*
Stemming Porter	*book, open, jason, awaken, school, bu, unabl, rememb, past*

Pada seluruh 4.446 sinopsis, rata-rata token per dokumen turun dari 375,6 (setelah tokenisasi) menjadi 184,5 (setelah *stopword* dan token pendek dibuang); kenaikan kecil dari 368,5 pada tahap awal ke 375,6 terjadi karena apostrof memecah satu kata menjadi dua token (misalnya *Arthur's* menjadi *arthur* dan *s*). *Stemming* menurunkan kosakata unik dari 53.557 menjadi 37.798 (turun 29,4%).

Untuk BERT, sinopsis ditokenisasi menggunakan WordPiece tanpa prapemrosesan manual. Panjang masukan dibatasi 256 atau 512 token, termasuk [CLS] dan [SEP]. Pada Tabel 3, seluruh sampel terpotong pada konfigurasi 256 token, sedangkan hanya sampel 1 dan 2 terpotong pada konfigurasi 512 token. Secara keseluruhan, 50,0% sinopsis melebihi 256 token dan 28,4% melebihi 512 token, dengan median 256 token. Konfigurasi 512 memperluas cakupan konteks, tetapi sebagian sinopsis tetap terpotong.

 ISSN:2302-8149
e-ISSN:2540-9719

Tabel 3 Jumlah token pada tiga sampel sinopsis

Sampel	Sub-token isi sebelum token khusus	Total token termasuk [CLS] dan [SEP]	Sub-token isi terpotong pada BERT 256	Sub-token isi terpotong pada BERT 512
1	1.064	1.066	810	554
2	719	721	465	209
3	478	480	224	0

Setelah prapemrosesan, data dibagi dengan *stratified split* 70/15/15 menjadi 3.112 data latih, 667 data validasi, dan 667 data uji (Tabel 4), genre minoritas tetap bersisa sangat sedikit pada data uji; psychology 13 sampel, romance 12, sports 12, dan travel 15, sehingga hasil pada pembagian ini rentan terhadap variasi dan perlu dilengkapi validasi silang (Subbab 3.5).

Tabel 4 Pembagian data (*stratified* 70/15/15)

Bagian	Jumlah Sampel	Catatan
Latih	3.112	Digunakan untuk melatih seluruh model
Validasi	667	Digunakan hanya oleh BERT untuk early stopping
Uji	667	Digunakan untuk evaluasi akhir pada pembagian tetap

4.3 Hasil pada Pembagian Tetap

Tabel 5 merangkum hasil ketigabelas konfigurasi pada 667 data uji dari pembagian tetap, mencakup akurasi, F1-macro, kappa, dan waktu latih (cetak tebal = tertinggi), dibandingkan dengan angka yang dilaporkan penelitian terdahulu [4].

Tabel 5 Hasil pembagian data tetap

Kode	Model	Akurasi (%)	F1-macro	Kappa	Waktu latih (detik)
[4]	RF 500 fitur, sebelum penyeimbangan	42,11	0,360	0,310	-
[4]	RF 500 fitur, sesudah penyeimbangan	46,67	0,443	0,387	-
A1	RF 500, resampling	45,88	0,422	0,388	2,0
A2	RF 500, tanpa penanganan	60,42	0,424	0,520	11,2
A3	RF 500, class_weight	61,77	0,586	0,553	8,1
A4	RF penuh, tanpa penanganan	61,47	0,426	0,530	28,7
A5	RF penuh, class_weight	65,97	0,646	0,598	25,1
B1	LR 500, class_weight	64,92	0,612	0,593	0,5
B2	SVM 500, class_weight	62,82	0,597	0,565	0,2
B3	LR penuh, class_weight	74,96	0,745	0,703	8,3
B4	SVM penuh, class_weight	76,01	0,769	0,717	0,3
C1	BERT 256, berbobot	79,91	0,841	0,763	852,5
C2	BERT 256, tanpa bobot	79,76	0,835	0,762	1.108,9
C3	BERT 512, berbobot	**80,96**	**0,842**	**0,776**	1.943,1
C4	BERT 512, tanpa bobot	**80,96**	0,825	**0,776**	1.934,4

Adaptasi prosedur *random resampling* penelitian terdahulu (A1) menghasilkan akurasi 45,88% dan F1-macro 0,422, mendekati hasil yang dilaporkan penelitian tersebut, yaitu akurasi 46,67% dan F1-macro 0,443. Namun, perbedaan data dan skema pembagian membatasi perbandingan langsung. Pada pembagian tetap penelitian ini, A1 tidak memperbaiki F1-macro dibandingkan kondisi tanpa penanganan (A2), yaitu 0,422 berbanding 0,424, dan menurunkan akurasi sebesar 14,54 poin persentase, dari 60,42% menjadi 45,88%. Pengambilan 100 sampel per genre mengurangi jumlah baris data latih dari 3.112 menjadi 1.000, atau sekitar 68%. Pengurangan data latih dan pengulangan sampel genre kecil diduga berkontribusi terhadap penurunan kinerja, tetapi pengaruh masing-masing faktor belum diuji secara terpisah. Sebaliknya, pembobotan kelas (A3) meningkatkan F1-macro dari 0,424 menjadi 0,586 (+0,162) tanpa membuang data latih, termasuk meningkatkan F1 *psychology* dan *sports* dari masing-masing 0,00 menjadi 0,46 dan 0,69.

Pada LR dan SVM dalam eksperimen ini, perubahan dari TF-IDF 500 fitur ke konfigurasi tanpa batas fitur meningkatkan F1-macro masing-masing dari 0,612 menjadi 0,745 dan dari 0,597 menjadi

 ISSN:2302-8149
e-ISSN:2540-9719

0,769. Peningkatan tersebut lebih besar daripada selisih F1-macro antara kedua algoritma pada representasi yang sama. Random Forest dengan representasi dan pembobotan yang sama (A5) mencapai F1-macro 0,646, tetapi tetap di bawah SVM. Pada pembagian tetap, seluruh konfigurasi BERT memperoleh akurasi dan F1-macro lebih tinggi daripada seluruh konfigurasi model klasik yang diuji. Pada 256 token, pembobotan kelas hanya memberi peningkatan tipis (C1: akurasi 79,91%; F1-macro 0,841 dibanding C2: 79,76%; 0,835). Pada 512 token, akurasi C3 dan C4 sama (80,96%) dan pembobotan hanya menaikkan F1-macro dari 0,825 (C4) menjadi 0,842 (C3). Perpanjangan token dari 256 ke 512 hampir tidak mengubah F1-macro pada model berbobot (C1 ke C3: 0,841 ke 0,842) dan sedikit menurunkannya tanpa bobot (C2 ke C4: 0,835 ke 0,825). Dengan demikian, keempat varian BERT berada pada kisaran akurasi 79,8-81,0% dan F1-macro 0,825-0,842 sehingga perbedaan antar-variannya kecil; C3 memiliki F1-macro tertinggi (0,842), yaitu 4,95 poin akurasi dan 0,073 F1-macro di atas SVM (B4), meskipun waktu latihnya jauh lebih lama (1.943,1 detik berbanding 0,3 detik). Karena kelas minoritas hanya memiliki 12-15 sampel uji, F1 1,00 pada sports dan 0,87 pada psychology perlu ditafsirkan hati-hati.

4.4 Hasil Validasi Silang

Tabel 6 menyajikan hasil validasi silang (cetak tebal = rata-rata tertinggi). Secara deskriptif, berdasarkan rata-rata F1-macro pada ketiga nilai k, BERT memperoleh skor tertinggi (0,77-0,82), diikuti SVM dan Regresi Logistik dengan TF-IDF tanpa batas fitur (0,73-0,77), RF TF-IDF tanpa batas fitur berbobot (0,63-0,64), model 500 fitur berbobot (0,52-0,58), serta RF tanpa pembobotan dan A1 (0,41-0,44). Simpangan baku menggambarkan variasi kinerja antar-*fold* dan pada banyak konfigurasi meningkat ketika k lebih besar.

Tabel 6 Hasil validasi silang (rata-rata ± simpangan baku antar-*fold*)

Kode	Model	k=3 Akurasi (%)	k=3 F1-macro	k=5 Akurasi (%)	k=5 F1-macro	k=10 Akurasi (%)	k=10 F1-macro
A1	RF 500, resampling	45,82 ± 2,19	0,423 ± 0,014	45,75 ± 1,57	0,428 ± 0,016	45,48 ± 2,54	0,419 ± 0,021
A2	RF 500, tanpa penanganan	57,78 ± 0,95	0,418 ± 0,007	58,37 ± 2,14	0,428 ± 0,031	59,04 ± 3,35	0,442 ± 0,024
A3	RF 500, class_weight	56,39 ± 0,96	0,521 ± 0,007	57,26 ± 1,99	0,540 ± 0,016	57,38 ± 2,63	0,536 ± 0,027
A4	RF penuh, tanpa penanganan	60,28 ± 0,72	0,412 ± 0,027	60,35 ± 1,95	0,408 ± 0,020	61,25 ± 2,46	0,429 ± 0,025
A5	RF penuh, class_weight	63,47 ± 0,65	0,631 ± 0,017	63,34 ± 1,16	0,627 ± 0,011	64,37 ± 2,27	0,636 ± 0,029
B1	LR 500, class_weight	61,00 ± 1,14	0,577 ± 0,017	61,07 ± 2,51	0,578 ± 0,016	61,09 ± 3,47	0,577 ± 0,040
B2	SVM 500, class_weight	57,67 ± 1,08	0,533 ± 0,011	59,49 ± 1,55	0,557 ± 0,018	59,70 ± 3,13	0,562 ± 0,034
B3	LR penuh, class_weight	74,72 ± 0,44	0,734 ± 0,009	75,37 ± 1,53	0,749 ± 0,023	75,98 ± 2,19	0,758 ± 0,031
B4	SVM penuh, class_weight	75,26 ± 0,38	0,755 ± 0,010	75,15 ± 1,81	0,752 ± 0,020	76,14 ± 2,59	0,766 ± 0,024
C1	BERT 256, berbobot	75,51 ± 1,11	0,781 ± 0,013	76,16 ± 1,23	0,794 ± 0,017	77,15 ± 2,06	0,801 ± 0,029
C2	BERT 256, tanpa bobot	76,07 ± 0,51	0,771 ± 0,014	76,61 ± 0,76	0,778 ± 0,014	77,22 ± 2,31	0,801 ± 0,025
C3	BERT 512, berbobot	77,26 ± 0,64	**0,809 ± 0,006**	**78,56 ± 1,09**	**0,816 ± 0,017**	78,65 ± 2,24	**0,816 ± 0,029**
C4	BERT 512, tanpa bobot	**78,27 ± 0,83**	0,797 ± 0,027	78,14 ± 0,66	0,799 ± 0,019	**78,81 ± 1,67**	0,807 ± 0,016

BERT dibanding SVM. Keempat varian BERT memperoleh rata-rata F1-macro lebih tinggi daripada SVM TF-IDF tanpa batas fitur (B4) pada seluruh nilai k, dengan selisih 0,016-0,064. Pada k=3 dan k=5, setiap varian memperoleh F1-macro lebih tinggi daripada B4 di semua *fold*, sedangkan pada k=10, C1-C3 lebih tinggi di 9 dari 10 *fold* dan C4 lebih tinggi di seluruh *fold*. Selisih terbesar terhadap B4 diperoleh C3, yaitu 0,049-0,064 F1-macro dan 2,00-3,42 poin akurasi. C3 juga memiliki rata-rata F1-macro tertinggi di antara varian BERT, dengan selisih terhadap C1 sebesar 0,014-0,028 dan terhadap C4 sebesar 0,008-0,017. Selisih ini dibaca sebagai kecenderungan pada rata-rata antar-fold, dengan perbedaan antar-varian BERT yang relatif kecil.

 ISSN:2302-8149
e-ISSN:2540-9719

Ablasi BERT: pembobotan kelas dan panjang token. Untuk menjawab pertanyaan penelitian keempat, pengaruh kedua faktor dianalisis melalui rata-rata F1-macro pada Tabel 6 dan F1 per genre pada Tabel 8. Pembobotan kelas cenderung menaikkan rata-rata F1-macro, tetapi peningkatannya relatif kecil: C1 di atas C2 sebesar 0,010, 0,016, dan 0,0005 pada k=3, 5, dan 10, sedangkan C3 di atas C4 sebesar 0,011, 0,017, dan 0,008. Peningkatan tersebut lebih kecil daripada peningkatan pada Random Forest, yang berkisar 0,09-0,22. Perpanjangan token dari 256 ke 512 juga menaikkan rata-rata F1-macro (C1 ke C3: +0,028, +0,022, +0,014; C2 ke C4: +0,027, +0,020, +0,007). Pada genre minoritas (Tabel 8), pengaruh keduanya tidak seragam: pembobotan pada 512 token menaikkan F1 romance (+0,034), sports (+0,027), dan travel (+0,050), tetapi menurunkan psychology (-0,024), sedangkan pada 256 token menurunkan romance (-0,010) dan psychology (-0,033). Perpanjangan token menaikkan F1 travel pada model berbobot (+0,035), tetapi menurunkan romance pada model tanpa bobot (-0,049).

Random Forest dengan fitur setara. Diberi TF-IDF tanpa batas fitur yang sama dengan SVM dan Regresi Logistik, pembobotan kelas jauh lebih efektif pada fitur penuh: F1-macro naik sekitar +0,21-0,22 (dari 0,408-0,429 ke 0,627-0,636 di k=3,5,10), dibanding hanya +0,09-0,11 pada 500 fitur, dengan akurasi naik ~3 poin (padahal turun 1,1-1,7 poin pada 500 fitur). Namun Random Forest tetap tertinggal sekitar 11,8 poin akurasi dan 0,12-0,13 F1-macro dari SVM pada fitur yang sama, dengan waktu rata-rata 15,2-29,7 detik per *fold* (Tabel 7), sekitar 49-90 kali waktu SVM.

***Random resampling* pada validasi silang.** Pola pada pembagian tetap juga ditemukan pada validasi silang: A1 tidak memberikan peningkatan F1-macro yang konsisten dibanding A2. F1-macro A1 berkisar 0,419-0,428, sedangkan A2 berkisar 0,418-0,442; A1 sedikit lebih tinggi pada k=3 (+0,005), setara pada k=5, dan lebih rendah pada k=10 (-0,022). Namun, akurasi A1 jauh lebih rendah, yaitu 45,48-45,82% dibanding 57,78-59,04%. Dengan demikian, *random resampling* mengorbankan akurasi tanpa memberikan keuntungan F1-macro yang konsisten pada skema evaluasi yang diuji.

Waktu komputasi. Untuk menjawab pertanyaan penelitian ketiga, Tabel 7 menyajikan perbandingan waktu komputasi seluruh konfigurasi. Model klasik membutuhkan 0,13-33,81 detik per *fold*, sedangkan BERT 662-1.954 detik (Tabel 7, rata-rata ± simpangan baku), dengan 512 token sekitar 1,6-2,4 kali lebih lama daripada 256 token. Waktu C3 sekitar 5.300-6.600 kali SVM (B4) dan 66-108 kali RF TF-IDF tanpa batas fitur berbobot (A5). Waktu k=10 diukur pada platform berbeda (Subbab 3.6).

Tabel 7 Waktu pelatihan per fold pada validasi silang (detik, rata-rata ± simpangan baku)

Kode	Model	k=3	k=5	k=10
A1	RF 500, resampling	2,16 ± 0,46	2,16 ± 0,42	1,27 ± 0,03
A2	RF 500, tanpa penanganan	9,43 ± 1,35	11,92 ± 0,42	6,10 ± 0,06
A3	RF 500, class_weight	8,24 ± 0,03	10,02 ± 0,52	5,07 ± 0,05
A4	RF penuh, tanpa penanganan	25,44 ± 1,33	33,81 ± 1,26	17,59 ± 0,44
A5	RF penuh, class_weight	21,26 ± 1,67	29,65 ± 1,68	15,20 ± 0,51
B1	LR 500, class_weight	0,24 ± 0,01	0,29 ± 0,02	0,30 ± 0,03
B2	SVM 500, class_weight	0,13 ± 0,00	0,17 ± 0,03	0,16 ± 0,00
B3	LR penuh, class_weight	9,75 ± 2,68	13,97 ± 2,15	14,66 ± 1,67
B4	SVM penuh, class_weight	0,24 ± 0,01	0,33 ± 0,06	0,31 ± 0,01
C1	BERT 256, berbobot	662,4 ± 86,9	875,4 ± 147,4	823,1 ± 3,5
C2	BERT 256, tanpa bobot	913,4 ± 35,8	882,7 ± 125,6	741,6 ± 47,7
C3	BERT 512, berbobot	1.575,9 ± 13,2	1.953,8 ± 105,1	1.646,4 ± 0,8
C4	BERT 512, tanpa bobot	1.458,5 ± 120,1	1.675,7 ± 156,6	1.682,5 ± 147,0

4.5 Analisis per Genre dan Pola Kesalahan

Tabel 8 menyajikan F1 per genre pada prediksi *out-of-fold* untuk k=10, mencakup model klasik terpilih dan keempat varian BERT.

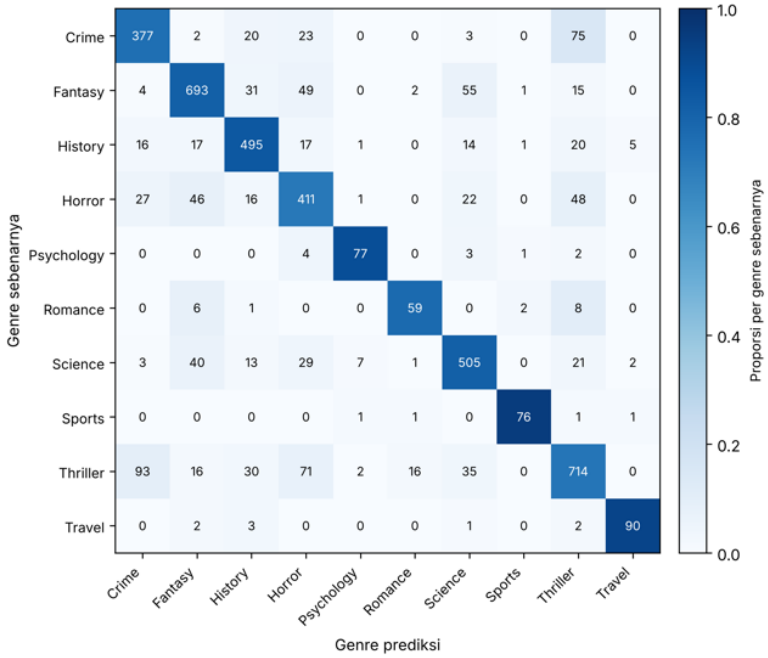
Tabel 8 F1 per genre pada prediksi *out-of-fold* (k=10)

Genre	Data	A2	A4	A5	B4	C1	C2	C3	C4
Crime	500	0,562	0,566	0,624	0,735	0,701	0,696	0,739	0,739
Fantasy	850	0,666	0,675	0,686	0,797	0,812	0,814	0,829	0,830

 ISSN:2302-8149
e-ISSN:2540-9719

Genre	Data	A2	A4	A5	B4	C1	C2	C3	C4
History	586	0,647	0,636	0,667	0,775	0,810	0,811	0,828	0,831
Horror	571	0,465	0,464	0,564	0,701	0,675	0,662	0,700	0,698
Psychology	87	0,109	0,129	0,829	0,861	0,865	0,898	0,875	0,899
Romance	76	0,120	0,000	0,230	0,595	0,766	0,776	0,761	0,727
Science	621	0,669	0,720	0,723	0,786	0,803	0,806	0,802	0,813
Sports	80	0,286	0,364	0,857	0,870	0,938	0,930	0,944	0,917
Thriller	977	0,580	0,624	0,568	0,742	0,754	0,761	0,758	0,765
Travel	98	0,356	0,167	0,646	0,819	0,883	0,870	0,918	0,868

Tanpa pembobotan, Random Forest gagal pada genre minoritas bahkan dengan TF-IDF tanpa batas fitur (F1 romance 0,000 dan travel 0,167 pada A4), sama dengan kesalahan pada psychology, sports, dan romance yang dilaporkan penelitian terdahulu [4]. Pembobotan kelas memulihkan kinerja psychology (0,829) dan sports (0,857), tetapi romance tetap lemah (0,230). Romance juga paling sulit bagi SVM (F1 0,595), sedangkan BERT C3 mencapai 0,761; selisih 0,166 merupakan yang terbesar antara BERT dan SVM pada seluruh genre, meskipun perlu ditafsirkan hati-hati karena romance hanya memiliki 76 sinopsis, atau sekitar 7-8 data uji per *fold* pada k=10. BERT C3 lebih tinggi daripada SVM B4 pada sembilan dari sepuluh genre dan hanya sedikit lebih rendah pada horror (0,700 berbanding 0,701). Keunggulan C3 tidak khusus pada genre minoritas: selisih terbesar setelah romance ada pada travel (+0,099) dan sports (+0,074), tetapi juga pada history (+0,053) dan fantasy (+0,032), sedangkan selisih F1 pada psychology (+0,014), crime (+0,004), dan horror (-0,001) relatif kecil. Di antara keempat varian BERT, tidak ada yang unggul pada semua genre minoritas: F1 romance berkisar 0,727-0,776 dan travel 0,868-0,918.



Gambar 3 Confusion matrix BERT C3 dari prediksi *out-of-fold* (k=10)

Pada prediksi *out-of-fold* BERT (C3) k=10 (Gambar 3), 949 dari 4.446 sinopsis (21,3%) salah diprediksi, terbanyak di antara genre besar yang berdekatan tema: thriller-crime (93 dan 75), thriller-horror (71), serta fantasy-science (55) dan fantasy-horror (49). Hanya 39 kesalahan (4,1%) berasal dari genre minoritas (psychology, romance, sports, travel), dan tidak satu pun masuk sepuluh kesalahan teratas; dari 76 sinopsis romance misalnya, 59 (77,6%) diprediksi benar dan 8 salah ke thriller. Ini berbeda dari penelitian terdahulu [4], yang menempatkan kesalahan terutama pada genre minoritas, tampaknya terkait model dan penanganan yang dipakai, karena pola itu muncul pada Random Forest tanpa pembobotan tetapi tidak pada BERT, meski pengaruh audit data dan ukuran data uji tidak diuji terpisah. Kesalahan antar genre besar yang bertetangga tematik juga konsisten dengan 96 baris yang terlibat konflik label pada data asli dan batasan satu genre per buku yang diakui studi lain [2], sehingga pendekatan *multilabel* patut dipertimbangkan.

 ISSN:2302-8149
e-ISSN:2540-9719

4.6 Pembahasan

Random resampling (A1) tidak memberikan peningkatan F1-macro yang konsisten meskipun akurasinya mendekati hasil Mahanani dan Fibriani [4]. Pada pembagian tetap, prosedur ini mengurangi jumlah baris data latih dari 3.112 menjadi 1.000, atau sekitar 68%, sehingga pengurangan data latih merupakan salah satu kemungkinan penyebab tidak meningkatnya kinerja. Sebaliknya, pembobotan kelas (A3, A5) memperbaiki F1-macro tanpa membuang data, sejalan dengan Kunaefi dkk. [13] dan Widayanti dan Kasih [24], yang menunjukkan bahwa penanganan melalui fungsi loss atau bobot kelas dapat bekerja baik tanpa mengubah data. Perluasan fitur dari 500 ke TF-IDF tanpa batas fitur dengan *bigram* juga menaikkan F1-macro Regresi Logistik dan SVM lebih besar daripada pergantian algoritma. Namun karena konfigurasi fitur berbeda dalam beberapa hal sekaligus (jumlah fitur, *bigram*, pembobotan sublinear), peningkatan ini tidak dapat dikaitkan pada satu faktor dan perlu diuji pada prapemrosesan lain [20].

Pada dataset Kaggle yang sama, Adhikari [14] menilai rasio kelas terbesar terhadap terkecil (sekitar 1:10) belum cukup timpang untuk ditangani, meski Random Forest di sana memiliki akurasi tertinggi (67,81%) tetapi F1 hanya 0,541, di bawah Regresi Logistik (0,660). Penelitian ini menunjukkan pola serupa: pada 500 fitur, pembobotan kelas hanya menaikkan akurasi 1,35 poin (A2 60,42% menjadi A3 61,77%), tetapi F1-macro naik 0,162 (0,424 menjadi 0,586). Dengan rasio kelas 12,9:1 setelah pembersihan, ketidakseimbangan tidak dapat diabaikan dan akurasi saja dapat menyesatkan. Akurasi Regresi Logistik dengan TF-IDF tanpa batas fitur di sini (74,7-76,0%) lebih tinggi daripada pada [14] (67,06%), tetapi perbedaan pembagian data, fitur, dan audit data membuat keduanya tidak dapat dibandingkan langsung.

Dibanding SVM dengan TF-IDF tanpa batas fitur, BERT C3 memperoleh akurasi 2,00-3,42 poin dan F1-macro 0,049-0,064 lebih tinggi pada ketiga nilai k. Selisih F1 per genre terbesar terdapat pada romance, travel, dan sports, tetapi tidak terbatas pada genre minoritas karena juga tampak pada history dan fantasy, sedangkan pada horror, F1 kedua model hampir sama. Pola ini mengisyaratkan bahwa representasi kontekstual membantu pada sebagian genre dengan data terbatas, meskipun penyebab pastinya belum dapat dipisahkan dari perbedaan pra-latih dan fungsi loss. Perbedaan ini dibaca sebagai kecenderungan pada rata-rata antar-fold. Gambaran ini selaras dengan Lin dkk. [8] bahwa pembanding linear yang kuat sering kompetitif, dan dengan Benítez-Andrades dkk. [9] bahwa keunggulan BERT dapat kecil dibanding biaya komputasinya.

Pada k=10, BERT (C3) membutuhkan rata-rata 1.646,4 detik per *fold*, sekitar 108 kali lebih lama daripada Random Forest TF-IDF tanpa batas fitur berbobot kelas (15,20 detik) dan sekitar 5.300 kali lebih lama daripada SVM (0,31 detik; rentang 5.300-6.600 kali pada k=3, 5, 10). Rasio ini mengukur waktu pelatihan saja, dengan BERT di GPU dan model klasik di CPU, sehingga bukan perbandingan *end-to-end*. Peningkatan F1-macro BERT perlu ditimbang terhadap tambahan waktu ini, terutama bila efisiensi menjadi prioritas. BERT dapat dipertimbangkan bila kinerja genre tertentu lebih diutamakan, misalnya romance dengan F1 0,761 berbanding 0,595 pada SVM. Pemilihan model karena itu sebaiknya mengikuti prioritas aplikasi, bukan semata kinerja rata-rata.

5 Kesimpulan

Penelitian ini membandingkan Random Forest, Regresi Logistik, dan SVM linear dengan BERT untuk klasifikasi genre novel dari sinopsis pada 4.446 data bersih yang tidak seimbang, dengan validasi silang *stratified* k=3, 5, dan 10 sebagai hasil utama serta pembagian tetap 70/15/15 sebagai pelengkap. *Random resampling* yang diadaptasi dari penelitian terdahulu tidak meningkatkan F1-macro secara konsisten dan menurunkan akurasi Random Forest dari 57,78-59,04% menjadi 45,48-45,82% pada validasi silang. Dalam konfigurasi yang diuji, perubahan representasi TF-IDF memberikan peningkatan F1-macro yang besar pada Regresi Logistik dan SVM. Pada k=5, penggunaan TF-IDF tanpa batas fitur dengan *bigram* meningkatkan F1-macro Regresi Logistik dari 0,578 menjadi 0,749 dan SVM dari 0,557 menjadi 0,752. Random Forest dengan representasi yang sama dan pembobotan kelas tetap tertinggal sekitar 12 poin persentase akurasi dan 0,12-0,13 F1-macro dari SVM. Pada BERT, pembobotan kelas dan perpanjangan input dari 256 menjadi 512 token cenderung meningkatkan rata-rata F1-macro, dengan selisih yang relatif kecil. C3 (512 token, berbobot kelas) memperoleh rata-rata F1-macro tertinggi di antara konfigurasi yang diuji, yaitu 0,809-

0,816. Dibandingkan dengan SVM berbobot kelas yang menggunakan TF-IDF tanpa batas fitur, C3 memperoleh akurasi 2,00-3,42 poin persentase lebih tinggi (77,3-78,7% berbanding 75,1-76,1%) dan F1-macro 0,049-0,064 lebih tinggi (0,809-0,816 berbanding 0,752-0,766).

Bagi pengembang sistem rekomendasi bacaan, SVM linear dengan TF-IDF tanpa batas fitur merupakan pembanding yang kuat dan efisien yang sebaiknya selalu dijalankan sebelum memakai model transformer. BERT layak dipertimbangkan bila kinerja pada genre tertentu seperti romance penting dan sumber daya komputasi tersedia. Apa pun modelnya, evaluasi sebaiknya memakai F1-macro, analisis per genre, dan validasi silang, karena akurasi pada satu pembagian data dapat menyesatkan.

Penelitian ini memiliki beberapa keterbatasan. Dataset hanya satu dan berbahasa Inggris, setiap konfigurasi dijalankan dengan satu *seed*, eksperimen dijalankan pada dua platform (Google Colab dan Kaggle) sehingga waktu k=10 tidak sebanding langsung dengan k=3 dan k=5, waktu BERT mencakup evaluasi validasi tiap epoch yang tidak ada pada model klasik, hiperparameter tidak dituning secara sistematis, dan setiap novel hanya diberi satu genre. Penelitian lanjutan dapat menerapkan klasifikasi *multilabel*, menguji teknik penyeimbangan lain dan ensemble khusus data tidak seimbang seperti Balanced Random Forest, mengulang eksperimen dengan beberapa *seed*, serta memakai dataset sinopsis berbahasa Indonesia.

Referensi

[1] E. M. N. Subroto and M. Faisal, "*A Multi-Label Book Genre Classification: Comparison of Machine Learning Techniques and Problem Transformation Methods*," *Jurnal Ilmiah Kursor*, Vol. 13, No. 2, Dec. 2025, DOI: 10.21107/kursor.v13i2.389.

[2] X. Yang and Z. Zhang, "*Enhancing Book Genre Classification with BERT and InceptionV3: A Deep Learning Approach for Libraries*," *PeerJ Comput. SCI.*, Vol. 11, p. e2934, Jun. 2025, DOI: 10.7717/peerj-cs.2934.

[3] A. Sethy, A. K. Rout, A. Uriti, and S. P. Yalla, "*A Comprehensive Machine Learning Framework for Automated Book Genre Classifier*," *Revue d'Intelligence Artificielle*, Vol. 37, No. 3, pp. 745–751, Jun. 2023, DOI: 10.18280/ria.370323.

[4] P. Mahanani and C. Fibriani, "*Novel Genre Classification based on Synopsis using the Random Forest Algorithm*," *SISTEMASI*, Vol. 15, No. 1, p. 296, Jan. 2026, DOI: 10.32520/stmsi.v15i1.5815.

[5] K. Hamaguchi, Y. Matsuda, and J. Kuroiwa, "*Potential of Contextual Representations for Genre Classification of Modern Japanese Literary Works using BERT*," *Nonlinear Theory and Its Applications, IEICE*, Vol. 17, No. 3, pp. 805–821, 2026, DOI: 10.1587/nolta.17.805.

[6] E. C. Garrido-Merchan, R. Gozalo-Brizuela, and S. Gonzalez-Carvajal, "*Comparing BERT Against Traditional Machine Learning Models in Text Classification*," *Journal of Computational and Cognitive Engineering*, Vol. 2, No. 4, pp. 352–356, Apr. 2023, DOI: 10.47852/bonviewJCCE3202838.

[7] N. Aydın, O. A. Erdem, and A. Tekerek, "*Comparative Analysis of Traditional Machine Learning and Transformer-based Deep Learning Models for Text Classification*," *Politeknik Dergisi*, Vol. 28, No. 2, pp. 445–452, Mar. 2025, DOI: 10.2339/politeknik.1469530.

[8] Y.-C. Lin, S.-A. Chen, J.-J. Liu, and C.-J. Lin, "*Linear Classifier: An often-Forgotten baseline for Text Classification*," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Stroudsburg, PA, USA: Association for Computational Linguistics, Aug. 2023, pp. 1876–1888. DOI: 10.18653/v1/2023.acl-short.160.

[9] J. A. Benítez-Andrades, J.-M. Alija-Pérez, M.-E. Vidal, R. Pastor-Vargas, and M. T. García-Ordás, "*Traditional Machine Learning Models and Bidirectional Encoder Representations from Transformer (BERT)–based Automatic Classification of Tweets about Eating Disorders: Algorithm Development and Validation Study*," *JMIR Med. Inform.*, Vol. 10, No. 2, p. e34492, Feb. 2022, DOI: 10.2196/34492.

[10] S. Henning, W. Beluch, A. Fraser, and A. Friedrich, "*A Survey of Methods for Addressing Class Imbalance in Deep-Learning based Natural Language Processing*," in *Proceedings of the 17th Conference of the European Chapter of the Association for Computational*

Linguistics, Stroudsburg, PA, USA: Association for Computational Linguistics, 2023, pp. 523–540. DOI: 10.18653/v1/2023.eacl-main.38.

[11] Atharva Inamdar, "*Book Genre Prediction*," Kaggle. Accessed: Oct. 14, 2026. [Online]. Available: https://www.kaggle.com/datasets/athu1105/book-genre-prediction

[12] M. C. Hinojosa Lee, J. Braet, and J. Springael, "*Performance Metrics for Multilabel Emotion Classification: Comparing Micro, Macro, and Weighted F1-Scores*," *Applied Sciences*, Vol. 14, No. 21, p. 9863, Oct. 2024, DOI: 10.3390/app14219863.

[13] A. Kunaefi, Z. Abidin, and R. Kusumawati, "Klasifikasi Berita *Hoaks* Bahasa Indonesia menggunakan *IndoBERT Fine-Tuning* dengan Pendekatan *Focal Loss* pada Data tidak Seimbang," *JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)*, Vol. 10, No. 2, pp. 1706–1714, May 2025, DOI: 10.29100/jipi.v10i2.7811.

[14] J. Adhikari, "*Leveraging Book Genre Classification using Machine Learning*," *DESIDOC Journal of Library & Information Technology*, Vol. 45, No. 3, pp. 169–176, May 2025, DOI: 10.14429/djlit.20026.

[15] K. Lee *et al.*, "*Deduplicating Training Data Makes Language Models Better*," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2022, pp. 8424–8445. DOI: 10.18653/v1/2022.acl-long.577.

[16] V. Solovyev, M. Solnyshkina, A. Ten, and N. Prokopyev, "*A BERT-based Classification Model: The Case of Russian Fairy Tales*," *Journal of Language and Education*, Vol. 10, No. 4, pp. 98–111, Dec. 2024, DOI: 10.17323/jle.2024.24030.

[17] D. Arroyan, M. R. Handayani, K. Umam, and M. A. Ulinuha, "Evaluasi Kinerja *Random Forest, SVM,* dan *Transformer* untuk Klasifikasi Komentar Judi Online di Youtube," *JUSTIN (Jurnal Sistem dan Teknologi Informasi)*, Vol. 14, No. 1, pp. 60–69, 2026, DOI: 10.26418/justin.v14i1.94059.

[18] G. F. S. Medantoro and M. Muljono, "*Comparative Analysis of IndoBERT and Classic Machine Learning Models for Sentiment Classification of Education Policy on Social Media X*," *Journal of Applied Informatics and Computing*, Vol. 10, No. 1, pp. 548–557, Feb. 2026, DOI: 10.30871/jaic.v10i1.11723.

[19] J. Wang, J. Wei, F. Tian, and Y. Wei, "*A Comparative Study of Machine Learning Models for Sentiment Analysis of Transboundary Rivers News Media Articles*," *Soft comput.*, Vol. 28, No. 23–24, pp. 13331–13347, Dec. 2024, DOI: 10.1007/s00500-024-10357-2.

[20] M. Siino, I. Tinnirello, and M. La Cascia, "*Is Text Preprocessing Still Worth the Time? A Comparative Survey on the Influence of Popular Preprocessing Methods on Transformers and Traditional Classifiers*," *Inf. Syst.*, Vol. 121, p. 102342, Mar. 2024, DOI: 10.1016/j.is.2023.102342.

[21] S. U. Hassan, J. Ahamed, and K. Ahmad, "*Analytics of Machine Learning-based Algorithms for Text Classification*," *Sustainable Operations and Computers*, Vol. 3, pp. 238–248, Jan. 2022, DOI: 10.1016/j.susoc.2022.03.001.

[22] K. Ghosh, C. Bellinger, R. Corizzo, P. Branco, B. Krawczyk, and N. Japkowicz, "*The Class Imbalance Problem in Deep Learning*," *Mach. Learn.*, Vol. 113, No. 7, pp. 4845–4901, Jul. 2024, DOI: 10.1007/s10994-022-06268-8.

[23] S. F. Taskiran, B. Turkoglu, E. Kaya, and T. Asuroglu, "*A Comprehensive Evaluation of Oversampling Techniques for Enhancing Text Classification Performance*," *SCI. Rep.*, Vol. 15, No. 1, p. 21631, Jul. 2025, DOI: 10.1038/s41598-025-05791-7.

[24] R. Widayanti and F. C. Kasih, "*Comparative Analysis of Baseline IndoBERT, Class-Weighted IndoBERT, and SMOTE with Support Vector Machine for Handling Imbalanced Sentiment Classification in Indonesian*," *Jurnal Teknik Informatika (Jutif)*, Vol. 7, No. 3, pp. 2857–2875, Jun. 2026, DOI: 10.52436/1.jutif.2026.7.3.5692.

[25] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "*BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423.