

GENERALISASI TEMPORAL MODEL KLASIK DAN MODEL BERBASIS TRANSFORMER UNTUK KLASIFIKASI SENTIMEN TINGKAT TARGET PADA JUDUL BERITA PERBANKAN INDONESIA

Eka Ratri Noor Wulandari¹⁾, Salnan Ratih Asriningtias²⁾, Dwi Utari Surya³⁾, Harnan Malik Abdullah⁴⁾, Rachmad Andri Atmoko⁵⁾, Muhammad Rifqi Ardhian⁶⁾, Mochamad Surya Rafliansyah⁷⁾, Mahendra Dwi Purwanto⁸⁾

^{1,2,3,4,5)} Departemen Industri Kreatif dan Digital, Universitas Brawijaya, Malang, Indonesia
^{6,7,8)} PT CMLABS INDONESIA DIGITAL, Indonesia

E-mail : ¹⁾ekaratri@ub.ac.id, ²⁾salnan@ub.ac.id, ³⁾d.utarisurya@ub.ac.id, ⁴⁾harnan_malik@ub.ac.id, ⁵⁾ra.atmoko@ub.ac.id, rifqiardhian@gmail.com⁶⁾, raflisurya824@gmail.com⁷⁾, mahendradwipurwanto@gmail.com⁸⁾

ABSTRAK

Klasifikasi sentimen tingkat target (*target-level sentiment classification*) pada judul berita berbahasa Indonesia menghadapi tantangan berupa keterbatasan konteks, ketidakseimbangan kelas, dan perubahan narasi pemberitaan. Penelitian ini bertujuan mengevaluasi kinerja model klasik dan IndoBERT dalam mengklasifikasikan sentimen tingkat target pada judul berita mengenai BCA, BNI, BRI, Bank Mandiri, dan BSI. *Dataset gold standard* disusun dari 2.000 judul berita yang diterbitkan pada tahun 2023–2025 melalui anotasi manual dan adjudikasi (penyelesaian ketidaksepakatan antaranotator). Kesepakatan antaranotator (*inter-annotator agreement*) menghasilkan Cohen's kappa sebesar 0,884 untuk relevansi dan 0,772 untuk sentimen. Sebanyak 1.957 judul yang dinyatakan relevan diberi label negatif, netral, atau positif. Model klasik berupa *logistic regression* dan *linear support vector machine* (SVM) dengan representasi *term frequency-inverse document frequency* (TF-IDF) kata serta gabungan kata dan karakter. Evaluasi model klasik dilakukan melalui validasi silang *5-fold* berbasis grup (*group-aware cross-validation*), sedangkan perbandingan dengan IndoBERT dilakukan melalui pengujian temporal menggunakan data 2023–2024 sebagai *development set* (pelatihan dan validasi) dan 538 judul 2025 sebagai data uji temporal terkunci (*locked temporal test set*). Hasil pengujian temporal menunjukkan bahwa SVM dengan fitur gabungan kata dan karakter mencapai Macro-F1 sebesar 0,8278 dan akurasi sebesar 0,9015. IndoBERT memperoleh Macro-F1 tertinggi sebesar 0,8592 dan *balanced accuracy* sebesar 0,8753. Sentimen negatif menjadi kelas yang paling sulit diklasifikasikan, sementara kinerja model bervariasi antarbank. Hasil tersebut menunjukkan bahwa IndoBERT lebih baik dalam mengenali ketiga kelas secara seimbang, sedangkan SVM menghasilkan akurasi keseluruhan yang lebih tinggi pada data uji. Kedua pendekatan berpotensi mendukung penyaringan judul berita dalam pemantauan media perbankan.

Kata kunci : analisis sentimen, berita perbankan, evaluasi temporal, IndoBERT, ketidakseimbangan kelas, pemantauan media, SVM

ABSTRACT

Automatically identifying target-level sentiment in Indonesian banking-news headlines is challenging because of limited context, class imbalance, and evolving media narratives. This study proposes a reproducible framework for target-level sentiment classification of headlines concerning five major banks: BCA, BNI, BRI, Bank Mandiri, and BSI. A gold-standard dataset was built from 2,000 headlines published in 2023–2025 through annotation and adjudication (Cohen's kappa of 0.884 for relevance and 0.772 for sentiment), yielding 1,957 relevant headlines labeled negative, neutral, or positive. Logistic regression and linear SVM with word- and character-level TF-IDF were compared with fine-tuned IndoBERT using five-fold group-aware cross-validation and a chronological evaluation, with 2023–2024 headlines for development and 538 headlines from 2025 as a locked temporal test set. The word-character SVM achieved a temporal Macro-F1 of 0.8278 and accuracy of 0.9015. IndoBERT obtained the highest temporal Macro-F1 (0.8592) and balanced accuracy of 0.8753. Negative sentiment was the most difficult class, and performance varied across banks. IndoBERT offers more class-balanced recognition, whereas the SVM remains competitive and computationally practical for banking-media screening.

Keywords: *banking news, class imbalance, IndoBERT, media monitoring, sentiment analysis*

1.PENDAHULUAN

Kepercayaan publik merupakan faktor penting dalam menjaga reputasi institusi perbankan. Perkembangan media daring mempercepat penyebaran informasi dan pembentukan persepsi masyarakat terhadap bank. Diskusi daring yang bernada negatif diketahui berkaitan dengan fluktuasi harga saham bank [1], sedangkan sentimen media dikaitkan dengan stabilitas simpanan dan risiko sistemik perbankan [2]. Temuan tersebut menunjukkan perlunya pemantauan sentimen secara sistematis untuk memahami pemberitaan mengenai institusi perbankan di media daring.

Analisis sentimen keuangan merupakan pendekatan komputasional untuk mengidentifikasi opini dan polaritas sentimen dalam teks keuangan yang tidak terstruktur. Pendekatan ini digunakan untuk mendukung pemantauan reputasi, penilaian sentimen

investor, dan identifikasi risiko [3]. Namun, penelitian analisis sentimen pada sektor perbankan menggunakan sumber data, definisi sentimen, prosedur anotasi, arsitektur model, dan protokol evaluasi yang beragam [4]. Perbedaan tersebut menyulitkan perbandingan kinerja antarmodel serta penilaian kemampuan model dalam mengklasifikasikan berita yang diterbitkan setelah periode pelatihan.

Klasifikasi sentimen pada judul berita memiliki tantangan tersendiri karena informasi yang disajikan singkat dan sentimennya sering kali tersirat. Selain itu, satu judul dapat menyebutkan beberapa entitas dengan sentimen yang berbeda terhadap masing-masing entitas. Sinha dkk. [5] mengidentifikasi 2.847 kasus yang melibatkan beberapa entitas dalam 10.753 judul berita keuangan yang dianotasi oleh manusia. Penelitian menggunakan model

transformer juga menunjukkan bahwa identifikasi sentimen terhadap entitas tertentu lebih kompleks daripada penentuan satu label sentimen untuk keseluruhan teks [6]. Oleh karena itu, penelitian ini menggunakan sentimen tingkat target, yaitu sentimen yang ditujukan kepada bank yang menjadi fokus analisis dalam setiap judul berita.

Keterbatasan konteks pada judul berita juga dapat menimbulkan ketidakpastian dalam pemberian label, terutama ketika membedakan sentimen netral dengan sentimen positif atau negatif yang tidak dinyatakan secara tegas. Tantangan ini semakin besar apabila distribusi data didominasi oleh kelas netral. Penanganan ketidakseimbangan kelas dapat meningkatkan kemampuan model dalam mengenali kelas minoritas, tetapi efektivitasnya bergantung pada karakteristik data, representasi fitur, dan algoritma yang digunakan [7]. Pada kondisi tersebut, akurasi yang tinggi belum tentu menunjukkan kemampuan model dalam mengenali seluruh kelas dengan baik. Oleh sebab itu, evaluasi perlu dilengkapi dengan Macro-F1, akurasi seimbang, *Matthews correlation coefficient* (MCC), dan *recall* setiap kelas.

Representasi *term frequency-inverse document frequency* (TF-IDF) yang dipadukan dengan algoritma klasifikasi linear menawarkan pendekatan yang efisien secara komputasi dan relatif mudah diinterpretasikan. Sementara itu, model *transformer* menghasilkan representasi kontekstual yang dapat digunakan untuk memahami hubungan antarkata dalam teks. Pengembangan model hibrida juga menunjukkan potensi peningkatan kinerja klasifikasi dan interpretabilitas [8]. Namun, pendekatan yang lebih ringan secara komputasi tetap unggul dalam kecepatan pemrosesan, ukuran model, dan kemudahan

interpretasi [9]. Dengan demikian, perbandingan model perlu mempertimbangkan kemampuan mengenali setiap kelas serta kebutuhan komputasinya.

Dalam teks berbahasa Indonesia, keragaman bentuk kata dan penggunaan istilah bahasa Inggris menambah tantangan klasifikasi sentimen. Kinerja model pada suatu bahasa juga tidak dapat langsung diasumsikan berlaku pada bahasa lain [10]. Penelitian pada data berbahasa Indonesia menunjukkan manfaat representasi kontekstual dan pendekatan hibrida [11], IndoBERT yang dikombinasikan dengan *recurrent layer* dan *attention mechanism* [12], serta *transfer learning* yang disesuaikan dengan karakteristik bahasa [13]. Namun, keberhasilannya tetap dipengaruhi oleh ukuran data, kesesuaian domain, dan strategi penyesuaian model atau *fine-tuning*. Pembedaan kelas netral dan negatif serta analisis teks yang memuat beberapa aspek juga masih menjadi tantangan [14]. Oleh karena itu, kemampuan model yang dilatih pada korpus sentimen umum berbahasa Indonesia perlu dievaluasi kembali sebelum diterapkan pada judul berita perbankan.

Selain pemilihan model, keandalan label acuan merupakan aspek penting dalam pengembangan sistem klasifikasi sentimen. Keluaran model bahasa besar atau *large language model* (LLM) dapat bervariasi akibat perbedaan *prompt*, konfigurasi model, dan proses inferensi yang bersifat stokastik [15]. Penelitian tentang penggunaan LLM sebagai alat bantu anotasi menekankan perlunya validasi terhadap label buatan manusia [16]. Dataset judul berita keuangan yang dianotasi secara manual juga telah digunakan untuk membandingkan pendekatan berbasis aturan, model *transformer*, dan LLM *zero-shot* [17]. Hal tersebut menunjukkan pentingnya prosedur

anotasi yang jelas dan pemeriksaan kesepakatan antarpenilai untuk menghasilkan label acuan yang dapat diandalkan.

Keandalan model juga perlu ditinjau berdasarkan waktu publikasi berita. Perubahan topik dan penggunaan bahasa dapat menimbulkan *concept drift* (pergeseran konsep) yang memengaruhi kinerja model pada data baru [18]. Dalam berita keuangan, perubahan tersebut dapat berkaitan dengan perkembangan narasi ekonomi dan munculnya kosakata yang terkait dengan peristiwa tertentu [19], serta perubahan sentimen judul berita dari waktu ke waktu [20]. Oleh sebab itu, pengujian pada data dari periode setelah pelatihan diperlukan untuk menilai kemampuan generalisasi temporal model. Perbandingan model klasik dan *transformer* melalui evaluasi tersebut, khususnya untuk klasifikasi sentimen tingkat target pada judul berita perbankan berbahasa Indonesia, masih terbatas.

Berdasarkan permasalahan tersebut, penelitian ini bertujuan mengembangkan dan mengevaluasi alur kerja klasifikasi sentimen tingkat target pada judul berita mengenai BCA, BNI, BRI, Bank Mandiri, dan BSI yang diterbitkan pada Januari 2023–Desember 2025. Model klasik berbasis TF-IDF dibandingkan dengan IndoBERT menggunakan validasi silang berkelompok dan pengujian temporal. Pada pengujian temporal, data tahun 2023–2024 digunakan sebagai data pengembangan, sedangkan data tahun 2025 digunakan sebagai data uji. Kontribusi penelitian ini meliputi penyusunan dataset judul berita perbankan yang divalidasi melalui anotasi manual dan adjudikasi, perbandingan representasi TF-IDF kata dan gabungan kata-karakter dengan IndoBERT, serta penerapan evaluasi lintas waktu untuk menilai kemampuan model pada berita dari periode berikutnya. Evaluasi

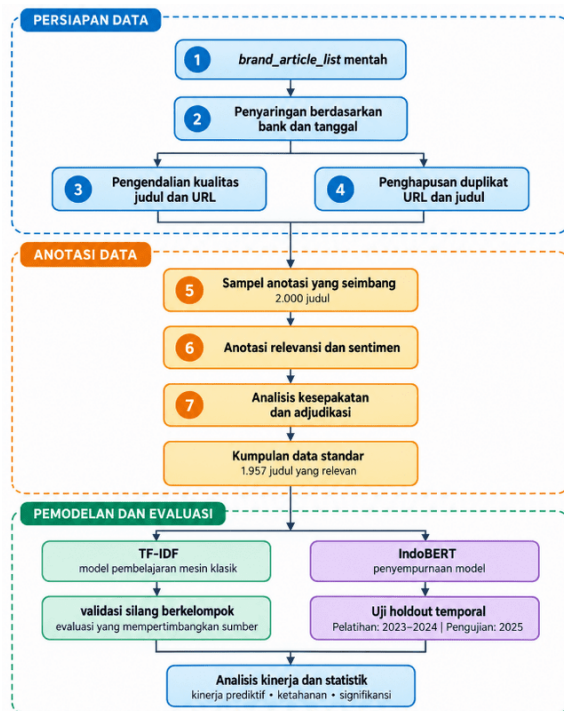
menggunakan Macro-F1 sebagai metrik utama, dilengkapi dengan metrik pendukung untuk menilai kinerja setiap kelas secara lebih menyeluruh.

2.METODE PENELITIAN

2.1 Sumber Data dan Cakupan Studi

Data diperoleh dari koleksi *brand_article_list* dalam basis data pemantauan media daring yang memuat ID artikel (*article identifier*), merek target (*brand_id*), media, judul, URL, dan tanggal publikasi. Cakupan penelitian meliputi BCA (*brand_id* 38), Bank Mandiri (39), BNI (40), BRI (41), dan BSI (43), dengan periode publikasi 1 Januari 2023–31 Desember 2025. Hanya judul yang digunakan sebagai masukan model, sedangkan URL dan metadata disimpan untuk ketertelusuran dan audit. Karena teks lengkap tidak tersedia secara konsisten, analisis dibatasi pada sentimen judul terhadap bank sasaran, sesuai dengan definisi tugas dan penggunaan data domain yang dianotasi oleh manusia [3].

Alur penelitian disajikan pada Gambar 1. Periode tiga tahun mendukung validasi silang berkelompok (*group-aware cross-validation*) pada seluruh data berlabel dan evaluasi kronologis menggunakan data 2023–2024 untuk pengembangan serta data 2025 untuk pengujian temporal.



Gambar 1. Alur kerja klasifikasi sentimen tingkat target yang diusulkan untuk judul berita perbankan Indonesia.

2.2 Penyaringan dan Deduplikasi

Pembersihan dilakukan sebelum anotasi dengan mengecualikan record yang judul atau URL-nya kosong, judul ternormalisasinya kurang dari 15 karakter, atau berasal dari halaman pencarian, tag, indeks, dan kumpulan artikel. Normalisasi judul mencakup perubahan ke huruf kecil, penghapusan spasi berlebih, akhiran nama penerbit, dan tanda elipsis di akhir judul yang terpotong. Normalisasi URL mencakup perubahan skema dan domain ke huruf kecil serta penghapusan awalan *www*, garis miring akhir, dan parameter pelacakan (*utm_source*, *gclid*, dan *fbclid*).

Deduplikasi dilakukan berdasarkan URL ternormalisasi dengan mempertahankan catatan paling awal, kemudian berdasarkan pasangan judul ternormalisasi dan *brand_id*.

Pengidentifikasi *title_group*, berupa *hash* SHA-1 judul ternormalisasi, digunakan dalam validasi silang untuk mencegah judul yang sama masuk ke data latih dan data uji sekaligus. Prosedur ini hanya mendeteksi kecocokan eksak (*exact match*), sehingga judul dengan redaksi berbeda tetapi bermakna sama (*near-duplicate*) dapat tetap lolos.

2.3 Definisi Sentimen Tingkat Target dan Protokol Anotasi

Unit analisis adalah pasangan judul h_i dan bank sasaran b_i , dengan fungsi klasifikasi pada Persamaan (1). Label mengacu pada bank yang ditunjukkan oleh *brand_id*, sehingga satu judul dapat memiliki sentimen berbeda terhadap bank yang berbeda.

$$f(h_i, b_i) \rightarrow y_i, \quad y_i \in \{\text{negatif, netral, positif}\} \quad (1)$$

Sentimen positif mencakup kinerja baik, pencapaian, pengakuan, peningkatan layanan, inovasi bermanfaat, atau keberhasilan korporat; netral mencakup informasi faktual tanpa evaluasi jelas; sedangkan negatif mencakup kegagalan, kerugian, kritik, pelanggaran, gangguan layanan, masalah hukum, insiden keamanan, atau ancaman reputasi bank sasaran. Judul ambigu atau kurang informatif diberi label tidak pasti.

Penelitian ini menggunakan tiga kategori sentimen terhadap bank sasaran, yaitu negatif, netral, dan positif. Kategori netral menunjukkan informasi tanpa evaluasi polaritas yang jelas, sedangkan keberadaan evaluasi positif dan negatif secara bersamaan terhadap bank yang sama belum dibedakan sebagai kategori tersendiri.

Sebanyak 2.000 judul dipilih melalui pengambilan sampel bertingkat (*stratified sampling*), masing-masing 400 judul untuk

setiap bank, untuk menjaga keterwakilan dalam dataset acuan (*gold standard*). Anotasi dilakukan secara hierarkis: penetapan relevansi (relevan, tidak relevan, atau tidak pasti), kemudian penetapan sentimen pada *record* yang relevan. Singkatan ambigu, penyebutan insidental, atau entitas serupa yang tidak terkait diberi label tidak relevan atau tidak pasti. Anotator hanya menggunakan judul tanpa menyimpulkan informasi dari URL, reputasi penerbit, atau pengetahuan eksternal. Protokol ini mengikuti prinsip pedoman anotasi eksplisit dan pengukuran kesepakatan antarpemilai [21], serta kebutuhan label yang sesuai domain (*in-domain*) untuk teks jurnalistik yang berbeda dari bahasa media sosial [22].

Kesepakatan relevansi dan sentimen dihitung terpisah menggunakan persentase kesepakatan (P_o) dan *Cohen's kappa* pada Persamaan (2), dengan P_e sebagai kesepakatan yang diharapkan secara kebetulan (*expected agreement by chance*) berdasarkan distribusi marginal label. Adjudikator meninjau judul, bank sasaran, kedua label awal, dan jenis ketidaksepakatan untuk menetapkan label akhir beserta alasan singkat. Label awal tetap disimpan untuk transparansi dan ketertelusuran.

$$\kappa = (P_o - P_e) / (1 - P_e) \quad (2)$$

2.4 Prapemrosesan Teks

Prapemrosesan dilakukan secara konservatif untuk mempertahankan informasi pada judul singkat, termasuk tanda baca, huruf kapital, singkatan, dan variasi morfologis. Pada model klasik, spasi berulang diganti satu spasi, judul kosong dibuang, dan pengubahan ke huruf kecil (*lowercasing*) dilakukan oleh *vectorizer* TF-IDF. Penghapusan *stopword*, *stemming*, dan penerjemahan tidak diterapkan agar istilah perbankan, negasi, singkatan, dan informasi

infleksi tetap terjaga. Pada IndoBERT, judul yang telah dibersihkan diproses menggunakan *tokenizer indobenchmark/indobert-base-pl* dengan panjang maksimum 128 token dan *padding* dinamis. Model pralatih khusus bahasa Indonesia dilaporkan lebih efektif menangkap fitur linguistik dibandingkan representasi generik [12].

2.5 Model Klasik: TF-IDF dan Pengklasifikasi Linear

Bobot TF-IDF didefinisikan pada Persamaan (3), dengan N adalah jumlah dokumen latih dan $df(t)$ adalah jumlah dokumen yang memuat *term* t . Representasi kata menggunakan *unigram* dan *bigram*, $min_df = 2$, $max_df = 98\%$, dan sublinear *term frequency scaling*, serta maksimum 60.000 fitur. Representasi karakter menggunakan *character n-gram* 3–5 dalam batas kata (*char_wb*), frekuensi dokumen minimum dua, dan maksimum 80.000 fitur untuk menangkap singkatan, variasi ejaan, afiks, nama produk, serta kata terpotong. Kedua representasi digabung melalui *feature union*.

$$tfidf(t, d) = tf(t, d) \cdot \log[(N + 1) / (df(t) + 1)] \quad (3)$$

Regresi logistik menggunakan $C = 2,0$ dan maksimum 3.000 iterasi, sedangkan SVM linear menggunakan $C = 1,0$. Keduanya menerapkan *balanced class weight* sesuai Persamaan (4), dengan K jumlah kelas dan N_c jumlah observasi latih pada kelas c

$$w_c = N / (K \cdot N_c) \quad (4)$$

2.6 Fine-Tuning IndoBERT

Model *indobenchmark/indobert-base-pl* disesuaikan (*fine-tuning*) dengan lapisan klasifikasi tiga kelas dan fungsi *softmax* pada representasi akhir token [CLS]. Mekanisme *self-attention* membentuk representasi setiap

token berdasarkan hubungannya dengan token lain sehingga model mempertimbangkan konteks judul, termasuk hubungan antara entitas, tindakan, dan peristiwa. Penyesuaian menggunakan judul berita perbankan beranotasi memungkinkan pembelajaran pola sentimen dalam domain penelitian. Prapemrosesan konservatif mempertahankan negasi dan istilah perbankan, sedangkan fungsi kerugian berbobot kelas mengurangi dominasi kelas netral. Masukan model berupa judul; bank sasaran menjadi acuan anotasi tanpa mekanisme atensi khusus target.

Data pengembangan mencakup judul relevan 2023–2024, dengan satu lipatan terstratifikasi berbasis kelompok sebagai validasi internal untuk pemilihan model dan penghentian dini (*early stopping*). Data 2025 tidak digunakan dalam pelatihan atau pemilihan hiperparameter. Hiperparameter dan fungsi kerugian disajikan pada Tabel 2 dan Persamaan (5).

$$\square = -(1/N) \sum w_y \log P(y_i | h_i) \quad (5)$$

2.7 Validasi dan Evaluasi

Model klasik dibandingkan menggunakan Stratified *Group K-Fold* dengan 5 *fold* untuk menjaga proporsi kelas dan menempatkan catatan dengan *title_group* sama dalam *fold* yang sama. *Vectorizer* hanya di-fit pada data latih setiap *fold* untuk mencegah kebocoran data (*data leakage*) dari statistik frekuensi dokumen. Setelah deduplikasi, seluruh 1.957 judul relevan memiliki *title_group* unik sehingga manfaat praktis dari penggunaan *group* ini terbatas. Evaluasi kronologis menggunakan data 2023–2024 untuk pengembangan dan data 2025 sebagai data uji temporal terkunci (*locked temporal test set*), yaitu tidak digunakan selama pengembangan model. Pemisahan kronologis digunakan untuk menilai kemampuan model pada berita

dari periode berikutnya, yang dapat memiliki perubahan topik, kosakata, dan komposisi sentimen. Protokol ini mengevaluasi generalisasi temporal, tetapi tidak secara khusus mengendalikan atau mengisolasi pengaruh kondisi politik dan ekonomi makro.

Macro-F1 menjadi metrik utama, dilengkapi akurasi seimbang, MCC, *Weighted-F1*, presisi dan *recall* per kelas, serta *confusion matrix*. Metrik agregat atau berbobot frekuensi dapat menutupi kegagalan pada kelas minoritas[23]. Model yang selalu memprediksi kelas mayoritas digunakan sebagai baseline.

Tabel 2. Hiperparameter *Fine-Tuning* IndoBERT.

Parameter	Nilai
Model prlatih	indobert-base-pl
Panjang maksimum	128 token
Laju pembelajaran	2×10^{-4}
Jumlah <i>epoch</i> maksimum	5

Parameter	Nilai
<i>Batch size</i> pelatihan / evaluasi	16 / 32
<i>Weight decay</i>	0,01
<i>Weight decay</i>	0,01
Pemanasan (<i>warm-up</i>)	10% langkah pembaruan
Patience penghentian dini	2 epoch
<i>random seed</i>	42
Fungsi kerugian	<i>Cross-entropy</i> berbobot kelas

3. HASIL DAN DISKUSI

3.1 Karakteristik Dataset dan Keandalan Anotasi

Dari 2.000 judul yang dianotasi, 1.957 relevan dan digunakan untuk klasifikasi tiga kelas, sedangkan 43 tidak relevan. Judul relevan mencakup 552 judul tahun 2023, 867 tahun 2024, dan 538 tahun 2025. Data 2023–2024 digunakan untuk pengembangan, sementara data 2025 menjadi data uji temporal independen (Tabel 3).

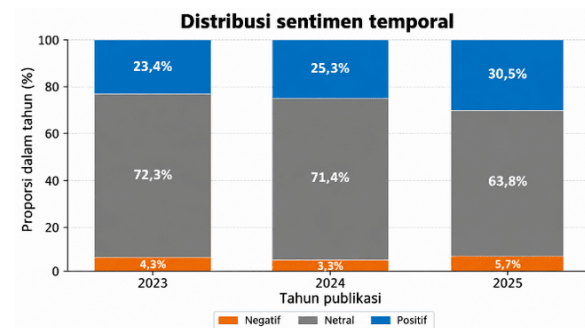
Kesepakatan langsung tercapai pada 1.804 judul, sedangkan 196 memerlukan adjudikasi. Kesepakatan relevansi mencapai 99,50% ($\kappa = 0,884$; hampir sempurna), sementara kesepakatan sentimen mencapai 90,47% ($\kappa = 0,772$; substansial). Perbedaan ini menunjukkan bahwa penentuan sentimen lebih sulit daripada relevansi, terutama pada judul promosi, aksi korporasi, insiden keamanan, dan informasi yang implikasinya terhadap bank tidak eksplisit.

Tabel 3. Karakteristik dataset dan keandalan anotasi.

Parameter	Nilai
Judul dianotasi / relevan / tidak relevan	2.000 / 1.957 / 43
Netral	1.361 (69,55%)
Positif	512 (26,16%)
Negatif	84 (4,29%)
Data pengembangan (2023–2024)	1.419
Data uji temporal (2025)	538
Kesepakatan langsung / adjudikasi	1.804 / 196
Kesepakatan relevansi; κ	99,50%; 0,884
Kesepakatan sentimen; κ	90,47%; 0,772

Distribusi kelas didominasi sentimen netral, dengan rasio netral terhadap negatif

sekitar 16,2:1. Dominasi ini selaras dengan sifat pemberitaan yang faktual dan informatif, tetapi memungkinkan akurasi tinggi meskipun pengenalan kelas negatif rendah. Proporsi sentimen positif dan negatif juga bervariasi antartahun (Gambar 2), sehingga komposisi kelas data uji temporal berbeda dari data pengembangan.



Gambar 2. Distribusi sentimen judul berita tahun 2023–2025 (persentase per tahun).

3.2 Kinerja Validasi Silang

Model mayoritas mencapai akurasi 0,6954, tetapi Macro-F1 hanya 0,2734, menunjukkan keterbatasan akurasi dalam menilai kinerja pada data yang tidak seimbang (Tabel 4). SVM dengan TF-IDF kata dan karakter menghasilkan kinerja terbaik di antara model klasik, dengan rata-rata Macro-F1 0,7885, melampaui SVM kata (0,7642) dan kata *logistic regression* (0,7585). Penambahan fitur karakter meningkatkan Macro-F1 sebesar 0,0243, yang mengindikasikan bahwa fitur karakter memberikan informasi tambahan dari singkatan, nama produk, variasi morfologis, dan kata terpotong.

Tabel 3. Kinerja validasi silang berkelompok lima lipatan.

Model	Akurasi	Akurasi seimbang	Macro-F1	Weighted-F1	MCC
Baseline mayoritas	0,6954	0,3333	0,2734	0,5706	0,0000
Kata + logistic regression	0,8646	0,7390	0,7585	0,8622	0,6894
Kata + SVM	0,8820	0,7218	0,7642	0,8767	0,7237
Kata-karakter + SVM	0,8983	0,7436	0,7885	0,8934	0,7630

3.3 Kinerja pada Set Uji Temporal

Pada 538 judul uji temporal tahun 2025, IndoBERT mencapai Macro-F1 tertinggi sebesar 0,8592 dan akurasi seimbang sebesar 0,8753, melampaui SVM kata-karakter masing-masing sebesar 0,0314 dan 0,0934 (Tabel 5). Hasil ini menunjukkan kemampuan pengenalan yang lebih seimbang antarkelas pada berita dari periode setelah pelatihan. Representasi kontekstual melalui *self-attention* memungkinkan model mempertimbangkan hubungan antartoken ketika kata berpolaritas tertentu tidak langsung menentukan sentimen terhadap bank. Sebagai contoh, penyebutan penipuan dapat merujuk pada pelanggaran oleh suatu pihak atau tindakan bank dalam mencegahnya. Penyesuaian pada domain perbankan dan pembobotan kelas juga berpotensi mendukung kinerja tersebut, sejalan dengan penelitian pada domain bahasa Indonesia[14],[22].

Namun, *self-attention* tidak menjamin ketepatan pengaitan sentimen dengan bank sasaran pada seluruh judul. Judul multientitas dapat menimbulkan ambiguitas, sedangkan informasi yang hilang pada judul terpotong tidak dapat dipulihkan hanya melalui hubungan antartoken. Karena

kontribusi mekanisme atensi belum diuji melalui ablasi atau evaluasi khusus judul multientitas, hasil ini menunjukkan kemampuan konfigurasi IndoBERT secara keseluruhan, bukan bukti bahwa permasalahan ambiguitas target telah sepenuhnya teratasi.

Tabel 4. Kinerja pada set uji temporal 2025 ($n = 538$).

Metrik	SVM kata-karakter	IndoBERT
Akurasi	0,9015	0,8866
Akurasi seimbang	0,7819	0,8753
Macro-F1	0,8278	0,8592
Weighted-F1	0,8979	0,8859
MCC	0,7970	0,7711

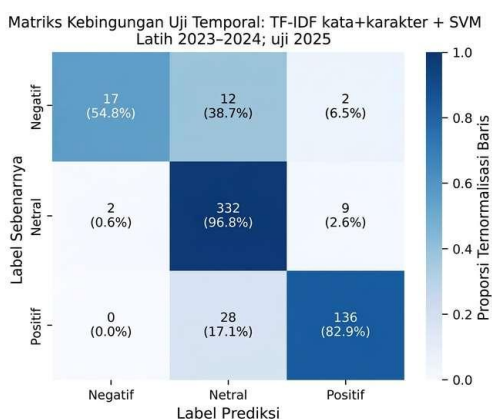
SVM unggul pada akurasi, *Weighted-F1*, dan MCC. Dari 343 judul netral (63,75% data uji), SVM mengenali 332 dengan benar (*recall* 0,9679), serta mencapai F1 kelas positif 0,8746. Kinerja ini mengindikasikan manfaat pola leksikal, seperti ungkapan peluncuran produk dan layanan, serta pola karakter berupa singkatan, nama produk, dan fragmen judul. Perbedaan peringkat mencerminkan karakteristik metrik: Macro-F1 menilai setiap kelas secara setara, sedangkan MCC merangkum korelasi keseluruhan antara prediksi dan label acuan. Dengan demikian, IndoBERT lebih seimbang antarkelas, sementara SVM menghasilkan lebih sedikit kesalahan total dan MCC lebih tinggi.

Kekuatan utama kedua model adalah kemampuannya mempertahankan kinerja klasifikasi pada data tahun 2025 yang tidak digunakan selama pengembangan.

IndoBERT mencapai Macro-F1 0,8592, sedangkan SVM kata-karakter mencapai 0,8278, menunjukkan bahwa pola yang dipelajari dari data 2023–2024 tetap berguna pada periode berikutnya. Temuan ini memberikan bukti generalisasi temporal dalam cakupan penelitian. Namun, perbedaan kinerja antar model belum dapat dikaitkan dengan ketahanan terhadap perubahan politik atau ekonomi makro karena variabel tersebut tidak dianalisis secara terpisah.

3.4 Kinerja Tingkat Kelas dan Analisis Kesalahan

SVM mencapai F1 sebesar 0,9287 pada kelas netral dan 0,8746 pada kelas positif, tetapi hanya 0,6800 pada kelas negatif. Presisi kelas negatif mencapai 0,8947, sedangkan *recall* hanya 0,5484. Dari 31 judul negatif, 17 diklasifikasikan dengan benar, 12 diprediksi netral, dan 2 positif (Gambar 3). Hasil ini menunjukkan kecenderungan konservatif: prediksi negatif umumnya benar, tetapi banyak judul negatif tidak terdeteksi. Sebanyak 28 dari 164 judul positif juga diprediksi netral.



Gambar 3. *Confusion matrix* ternormalisasi per baris dari SVM kata-karakter pada set uji temporal 2025.

SVM salah mengklasifikasikan 53 dari 538 judul (9,85%). Kesalahan meliputi

positif→netral (28), negatif→netral (12), netral→positif (9), netral→negatif (2), dan negatif→positif (2), tanpa kesalahan positif→negatif. Sebanyak 40 kesalahan (75,47%) merupakan kelas minoritas yang salah diprediksi sebagai netral.

Pemeriksaan kualitatif mengindikasikan empat faktor: keterbatasan contoh negatif (84 judul; 4,29%), penyampaian peristiwa negatif dengan gaya bahasa faktual, ambiguitas peran bank sebagai korban atau pihak yang merespons, serta hilangnya konteks pada judul terpotong. Sebagai contoh, judul “Serangan Siber ke Sistem BCA Tembus 4 Miliar pada 2024” memuat peristiwa negatif dalam struktur kalimat yang informatif yang dapat menyerupai berita netral. Aktivitas positif, seperti perluasan cabang, program komunitas, kemitraan, dan peluncuran produk, juga kerap diprediksi netral karena manfaat bagi bank tidak dinyatakan secara eksplisit.

3.5 Kinerja Tingkat Bank

Kinerja SVM kata-karakter bervariasi antarbank (Tabel 6). BNI mencapai Macro-F1 tertinggi (0,9360), diikuti BRI (0,9221), Bank Mandiri (0,8722), dan BSI (0,8602). BCA memperoleh Macro-F1 terendah (0,6178) meskipun akurasi 0,8500, serta kesalahan terbanyak (18 dari 120 judul; 15,00%). Perbedaan ini mengindikasikan pengenalan kelas minoritas yang kurang merata. Karena ukuran dan komposisi kelas berbeda antarbank, hasil ini digunakan sebagai analisis diagnostik per subkelompok (*subgroup diagnostics*), bukan peringkat definitif. Kinerja agregat yang tinggi belum menjamin keandalan seragam pada setiap bank.

Tabel 5. Kinerja SVM kata-karakter per bank pada set uji temporal 2025.

Bank	n	Akurasi	Balanced accuracy	Macro-F1	MCC	Jumlah Salah Prediksi
BNI	106	0,9151	0,9296	0,9360	0,8339	9
BRI	94	0,9362	0,8965	0,9221	0,8685	6
Bank Mandiri	105	0,9143	0,9268	0,8722	0,8230	9
BSI	113	0,9027	0,8142	0,8602	0,8215	11
BCA	120	0,8500	0,6120	0,6178	0,5938	18

3.6 Dampak Ketidakseimbangan Kelas dan Implikasi bagi Pemantauan Media

Dominasi kelas netral (69,55%) memungkinkan model mayoritas mencapai akurasi 0,6954, tetapi Macro-F1 hanya 0,2734 dan MCC nol. Hasil ini menegaskan perlunya Macro-F1 sebagai metrik utama, didukung akurasi seimbang, MCC, *Weighted-F1*, *recall* per kelas, dan *confusion matrix* untuk mengungkap kegagalan pada kelas minoritas [23].

Dalam pemantauan media perbankan, karena *recall* kelas negatif masih rendah, klasifikasi otomatis sebaiknya digunakan untuk membantu memprioritaskan human review, terutama pada berita berisiko tinggi. IndoBERT menunjukkan kinerja lebih seimbang antarkelas, sedangkan SVM dapat menjadi pilihan untuk penyaringan dengan kebutuhan komputasi lebih ringan. Alur hibrida dapat dikembangkan dengan SVM sebagai penyaring awal, lalu IndoBERT atau analisis meninjau judul yang berada dekat *decision boundary*, berkaitan dengan keamanan siber dan penipuan (*fraud*), atau menyebut banyak organisasi (*multi-entity*). Efektivitas alur hibrida ini masih perlu diuji.

Pemantauan operasional juga perlu mencakup *recall* per kelas, kinerja per bank, *temporal drift* (pergeseran temporal), dan jumlah prediksi berkeyakinan rendah (*low-confidence predictions*). Skema tiga kelas membatasi kemampuan sistem dalam merepresentasikan judul yang memuat evaluasi positif dan negatif terhadap bank yang sama. Karena kelas campuran belum dianotasi secara khusus, frekuensi dan kinerja pengenalannya belum dapat ditentukan. Dengan demikian, hasil penelitian menggambarkan kemampuan klasifikasi dalam skema negatif, netral, dan positif.

4. KESIMPULAN DAN SARAN

Penelitian ini mengembangkan dan mengevaluasi alur klasifikasi sentimen tingkat target menggunakan 2.000 judul beranotasi (1.957 relevan), validasi silang berbasis grup, dan data uji temporal 2025 yang terkunci. SVM kata-karakter menjadi model klasik terbaik dengan Macro-F1 validasi silang 0,7885 dan temporal 0,8278. IndoBERT mencapai Macro-F1 temporal tertinggi (0,8592) dan akurasi seimbang 0,8753, sedangkan SVM unggul pada akurasi (0,9015), *Weighted-F1*, dan MCC (0,7970). IndoBERT menunjukkan kinerja lebih seimbang antarkelas, sementara SVM menghasilkan lebih sedikit kesalahan keseluruhan pada distribusi uji yang tidak seimbang.

Tantangan utama adalah pengenalan sentimen negatif yang jarang dan sering tersirat, serta variasi kinerja antarbank. Evaluasi perlu mencakup metrik yang peka terhadap ketidakseimbangan kelas (*class-sensitive metrics*), *recall* per kelas, *confusion matrix*, dan kinerja per bank. Untuk penerapan operasional, klasifikasi dapat membantu penyaringan dan penentuan

prioritas, disertai peninjauan manusia pada prediksi berkeyakinan rendah, berita multientitas, serta kasus penipuan dan keamanan siber. Penelitian selanjutnya disarankan memperkaya contoh negatif, menguji model pada periode dan bank lain, serta mengevaluasi judul multientitas dengan masukan bank sasaran yang eksplisit. Uji ablasi diperlukan untuk membedakan kontribusi representasi kontekstual dan pembobotan kelas terhadap kinerja model. Temuan terbatas pada sentimen judul lima bank dan tidak merepresentasikan sentimen artikel lengkap, opini publik, bias media, kepuasan nasabah, atau reputasi korporat. Selanjutnya, penelitian dapat mengembangkan pedoman anotasi sentimen campuran dan mengevaluasi manfaatnya melalui anotasi ulang serta pengukuran kesepakatan antarpemilai.

DAFTAR PUSTAKA

- [1] A.-M. Casado-Molina, M.-M. Rojas-de-Gracia, P. Alarcón-Urbistondo, and C. M. Q. Ramos, "Banking reputation and its impact on stock markets: A big data analysis through online comments," *Int. J. Bank Mark.*, vol. 43, no. 5, pp. 979–998, 2025, doi: 10.1108/IJBM-07-2024-0419.
- [2] Y. Fang, Q. Wang, Y. Wang, and Y. Yuan, "Media sentiment, deposit stability and bank systemic risk: Evidence from China," *Int. Rev. Econ. Finance*, vol. 91, pp. 1150–1172, 2024, doi: 10.1016/j.iref.2024.01.067.
- [3] K. Du, F. Xing, R. Mao, and E. Cambria, "Financial sentiment analysis: Techniques and applications," *ACM Comput. Surv.*, vol. 56, no. 9, p. 220:1–220:42, 2024, doi: 10.1145/3649451.
- [4] M. Jayanthakumaran, N. Shukla, B. Pradhan, and G. Beydoun, "A systematic review of sentiment analytics in banking headlines," *Decis. Anal. J.*, vol. 15, p. 100584, 2025, doi: 10.1016/j.dajour.2025.100584.
- [5] A. Sinha, S. Kedas, R. Kumar, and P. Malo, "SEntFiN 1.0: Entity-aware sentiment analysis for financial news," *J. Assoc. Inf. Sci. Technol.*, vol. 73, no. 9, pp. 1314–1335, 2022, doi: 10.1002/asi.24634.
- [6] R. Pan, J. A. García-Díaz, F. García-Sánchez, and R. Valencia-García, "Evaluation of transformer models for financial targeted sentiment analysis in Spanish," *PeerJ Comput. Sci.*, vol. 9, p. e1377, 2023, doi: 10.7717/peerj-cs.1377.
- [7] S. F. Taşkıran, B. Türkoğlu, E. Kaya, and T. Aşuroğlu, "A comprehensive evaluation of oversampling techniques for enhancing text classification performance," *Sci. Rep.*, vol. 15, p. 21631, 2025, doi: 10.1038/s41598-025-05791-7.
- [8] M. A. Jahin, M. S. H. Shovon, M. F. Mridha, M. R. Islam, and Y. Watanobe, "A hybrid transformer and attention based recurrent neural network for robust and interpretable sentiment analysis of tweets," *Sci. Rep.*, vol. 14, p. 24882, 2024, doi: 10.1038/s41598-024-76079-5.
- [9] M. Rizinski, H. Peshov, K. Mishev, M. Jovanovik, and D. Trajanov, "Sentiment analysis in finance: From transformers back to eXplainable lexicons (XLex)," *IEEE Access*, vol. 12, pp. 7170–7198, 2024, doi: 10.1109/ACCESS.2024.3349970.
- [10] M. S. U. Miah, M. M. Kabir, T. B. Sarwar, M. Safran, S. Alfarhood, and M.

- F. Mridha, “A multimodal approach to cross-lingual sentiment analysis with ensemble of transformer and LLM,” *Sci. Rep.*, vol. 14, p. 9603, 2024, doi: 10.1038/s41598-024-60210-7.
- [11] C.-H. Lin and U. Nuha, “Sentiment analysis of Indonesian datasets based on a hybrid deep-learning strategy,” *J. Big Data*, vol. 10, p. 88, 2023, doi: 10.1186/s40537-023-00782-9.
- [12] H. Ahmadian, T. F. Abidin, H. Riza, and K. Muchtar, “Hybrid models for emotion classification and sentiment analysis in Indonesian language,” *Appl. Comput. Intell. Soft Comput.*, vol. 2024, p. 2826773, 2024, doi: 10.1155/2024/2826773.
- [13] C. Shaw, P. LaCasse, and L. Champagne, “Exploring emotion classification of Indonesian tweets using large scale transfer learning via IndoBERT,” *Soc. Netw. Anal. Min.*, vol. 15, p. 22, 2025, doi: 10.1007/s13278-025-01439-6.
- [14] R. I. Perwira, V. A. Permadi, D. I. Purnamasari, and R. P. Agusdin, “Domain-specific fine-tuning of IndoBERT for aspect-based sentiment analysis in Indonesian travel user-generated content,” *J. Inf. Syst. Eng. Bus. Intell.*, vol. 11, no. 1, pp. 30–40, 2025, doi: 10.20473/jisebi.11.1.30-40.
- [15] D. Herrera-Poyatos *et al.*, “An overview of model uncertainty and variability in LLM-based sentiment analysis: Challenges, mitigation strategies, and the role of explainability,” *Front. Artif. Intell.*, vol. 8, p. 1609097, 2025, doi: 10.3389/frai.2025.1609097.
- [16] M. Mathebula, A. Modupe, and V. Marivate, “ChatGPT as a text annotation tool to evaluate sentiment analysis on South African financial institutions,” *IEEE Access*, vol. 12, pp. 144017–144043, 2024, doi: 10.1109/ACCESS.2024.3464374.
- [17] I. Muhammad and M. Rospocher, “On assessing the performance of LLMs for target-level sentiment analysis in financial news headlines,” *Algorithms*, vol. 18, no. 1, p. 46, 2025, doi: 10.3390/a18010046.
- [18] G. Fenza, M. Gallo, V. Loia, A. Petrone, and C. Stanzone, “Concept-drift detection index based on fuzzy formal concept analysis for fake news classifiers,” *Technol. Forecast. Soc. Change*, vol. 194, p. 122640, 2023, doi: 10.1016/j.techfore.2023.122640.
- [19] I. A. Khandokar and P. Deshpande, “SSABE-TSCM: Drift-aware and interpretable financial sentiment analysis for low-resource Bangla via adaptive semi-supervised and temporal contrastive modeling,” *Front. Artif. Intell.*, vol. 9, p. 1724407, 2026, doi: 10.3389/frai.2026.1724407.
- [20] D. Rozado, R. Hughes, and J. Halberstadt, “Longitudinal analysis of sentiment and emotion in news media headlines using automated labelling with Transformer language models,” *PLOS ONE*, vol. 17, no. 10, p. e0276367, 2022, doi: 10.1371/journal.pone.0276367.
- [21] Md. E. Islam *et al.*, “SentiGOLD: A large Bangla gold standard multi-domain sentiment analysis dataset and its evaluation,” in *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, Association for Computing Machinery, 2023, pp. 4207–4218. doi: 10.1145/3580305.3599904.
- [22] D. M. Dama, T. Mardiana, R. Supriyadi, Z. P. Putra, D. Octaviano, and A.

Bayhaqy, “Fine-tuning IndoBERTa for Indonesian digital news sentiment classification,” *Artif. Intell. Appl.*, Jun. 2026, doi: 10.47852/bonviewAIA62027506.

- [23] M. C. Hinojosa Lee, J. Braet, and J. Springael, “Performance metrics for multilabel emotion classification: Comparing micro, macro, and weighted F1-scores,” *Appl. Sci.*, vol. 14, no. 21, p. 9863, 2024, doi: 10.3390/app14219863.