Breaking Through Noise with Fine-Tuned Efficiency: Hybrid Low-Parameter Models for Classifying Astra Cita's Topics

Evans Kizito
Departemen Matematika
Universitas Indonesia
Depok, Indonesia
evan.naibaho@gmail.com

Mohammad Raffy Zeidan
Departemen Matematika
Universitas Indonesia
Depok, Indonesia
raffy.zeidan@gmail.com

Muhammad Binar Raffi Lazuardi
Departemen Matematika
Universitas Indonesia
Depok, Indonesia
mohammadbinar6@gmail.com

***Abstract*—Analisis wacana publik terkait delapan pilar Astacita di media sosial X menjadi krusial untuk memahami aspirasi dan kritik masyarakat di era digital. Tugas klasifikasi teks ini menghadapi tantangan signifikan berupa data yang tidak seimbang (*imbalanced*) dan potensi label yang berisik (*noisy*), yang membatasi performa model-model standar. Untuk mengatasi hal ini, penelitian ini mengusulkan sebuah *pipeline machine learning komperhensif* yang berfokus pada pendekatan data-sentris. Metodologi ini dimulai dengan *Confident Learning* menggunakan prediksi *out-of-fold* (OOF) dari 5-Fold Cross-Validation untuk mengidentifikasi dan membersihkan data latih yang kemungkinan besar salah label. Dataset yang telah dibersihkan kemudian diperkaya menggunakan teknik *pseudo-labeling* untuk memanfaatkan informasi dari data tes secara aman. Model akhir dilatih menggunakan Stratified K-Fold Cross-Validation pada data gabungan ini dengan model dasar indolem/indobertweet-base-uncased yang telah dioptimalkan melalui *hyperparameter tuning*. Untuk memaksimalkan robustisitas, prediksi dari *ensemble* Stratified K-Fold ini digabungkan (*blended*) dengan prediksi dari model TF-IDF + *Logistic Regression*. Hasilnya, pendekatan berlapis ini berhasil mencapai skor *Balanced Accuracy* sebesar $0,613$. Penelitian ini menunjukkan bahwa dalam menghadapi data dunia nyata yang tidak sempurna, strategi pembersihan data dan *ensembling* hibrida yang *robust* lebih krusial daripada sekadar optimasi model tunggal.**

***Index Terms*—Klasifikasi Teks, *Natural Language Processing*, Astacita, *Confident Learning*, *Pseudo-Labeling*, *Stratified K-Fold Cross-Validation*, *Ensembling Hibrida*, IndoBERTweet.**

I. Pendahuluan

Di era digital saat ini, analisis wacana publik telah menjadi komponen krusial dalam memahami dinamika sosial-politik dan mengevaluasi arah kebijakan nasional. Visi Indonesia Emas 2045, yang dicanangkan sebagai panduan strategis jangka panjang, bertujuan untuk mentransformasi Indonesia menjadi negara maju yang berdaulat dan berkelanjutan [8]. Sebagai implementasi tahap awal dari visi tersebut, pemerintahan saat ini mengusung kerangka kerja delapan misi yang dikenal sebagai Astacita, yang mencakup berbagai sektor mulai dari penguatan ideologi hingga reformasi birokrasi. Platform media sosial seperti X (sebelumnya Twitter) telah berevolusi menjadi arena utama bagi wacana publik, di mana sentimen, aspirasi, dan kritik terhadap program seperti Astacita diekspresikan secara masif dan *real-time*. Dengan tingkat penetrasi internet di Indonesia yang mencapai 79.5% pada awal tahun 2024, analisis terhadap data dari platform ini menawarkan wawasan yang sangat berharga [2].

Namun, proses analisis ini menghadapi tantangan teknis yang signifikan. Data teks dari media sosial memiliki karakteristik unik: bersifat tidak terstruktur, informal, sangat singkat, dan sarat akan *noise* seperti *slang*, singkatan, dan sarkasme [6]. Selain itu, dataset yang digunakan dalam penelitian ini menunjukkan adanya ketidakseimbangan kelas yang parah antar label, yang berisiko menghasilkan model yang bias. Lebih lanjut, eksperimen awal mengindikasikan adanya potensi label yang tidak konsisten (*noisy labels*) yang dapat menyesatkan proses belajar model secara signifikan.

Penelitian ini bertujuan untuk merancang dan mengimplementasikan sebuah *pipeline machine learning end-to-end* yang *robust* untuk mengatasi tantangan tersebut. Tujuan utamanya adalah memaksimalkan skor metrik Balanced Accuracy dengan menerapkan strategi pemilihan model yang tepat, pembersihan data secara sistematis, dan metode pelatihan yang stabil. Ruang lingkup kerja ini terbatas pada dataset yang disediakan, dengan fokus pada *fine-tuning* model Transformer berbahasa Indonesia, `indolem/indobertweet-base-uncased`, yang secara spesifik dioptimalkan untuk teks media sosial.

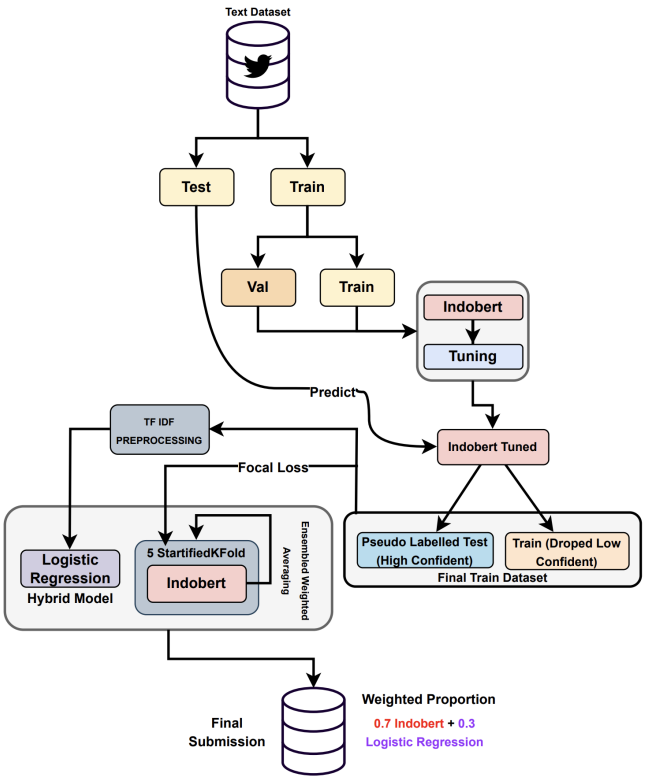
Makalah ini diawali dengan pemaparan metodologi yang mencakup pendekatan *Confident Learning* untuk pembersihan data, *Stratified K-Fold Cross-Validation* untuk pelatihan model, dan *Ensembling Hibrida* sebagai strategi prediksi akhir. Selanjutnya, hasil analisis dan perbandingan performa dari setiap tahap akan disajikan, dan diakhiri dengan rangkuman kesimpulan dari seluruh proses yang telah dilakukan.

II. Metodologi

A. Alur Kerja Penelitian

Penelitian ini mengadopsi alur kerja *machine learning* yang sistematis dan iteratif. Dimulai dari pemilihan model dasar, optimasi, pembersihan data, penambahan data, hingga metode *ensemble* hibrida untuk mencapai performa maksimal. Alur

kerja ini dirancang secara spesifik untuk mengatasi tantangan data teks media sosial yang *noisy* dan tidak seimbang.



Gambar 1. Research Flow Method

B. Dataset

Dataset penelitian diperoleh dari *ANFORCOM: Data Science Competition*, yang terdiri atas 5,000 data latih dan 5,000 data uji berasal dari platform X. Setiap sampel merupakan teks singkat yang telah diberi label ke dalam delapan kategori topik. Distribusi label pada data latih bersifat tidak seimbang, dengan rincian: *ideologi* (24.4%), *pertahanan* (21.42%), *reformasi* (15.42%), *harmoni* (13.28%), *sumber daya manusia (SDM)* (8.76%), *pekerjaan* (6.94%), *pemerataan* (6.12%), dan *hilirisasi* (3.64%). Ketidakseimbangan ini menjadi pertimbangan penting dalam pemilihan metode praproses dan pemodelan.

C. Pemisahan dan Persiapan Data

Dataset dibagi menjadi tiga bagian, yaitu data latih, data validasi, dan data uji untuk memastikan pengembangan dan evaluasi model berjalan secara andal. Data latih digunakan untuk melatih model IndoBERT maupun *Logistic Regression*, sementara data validasi dimanfaatkan dalam proses penyetelan *hyperparameter* dan pemilihan model terbaik. Adapun data uji memiliki dua fungsi utama, yaitu untuk mengevaluasi performa sistem akhir dan sebagai sumber tambahan data pada tahap *pseudo-labelling* di *pipeline* IndoBERT.

Terkait tahap praproses, strategi yang diterapkan berbeda antara kedua model. Pada IndoBERT, tidak dilakukan praproses tambahan karena model ini telah dilengkapi *tokenizer* internal yang secara efektif menangani teks mentah. Pendekatan ini dipilih untuk menjaga keutuhan semantik kalimat dan memaksimalkan representasi bahasa yang telah dimiliki IndoBERT.

Sebaliknya, pada *Logistic Regression* diperlukan praproses teks eksplisit. Setiap dokumen diubah menjadi representasi numerik menggunakan metode Term Frequency–Inverse Document Frequency (TF-IDF). Representasi ini menangkap tingkat kepentingan relatif sebuah kata dalam korpus sehingga memungkinkan *Logistic Regression* menangani data teks yang bersifat jarang dengan lebih efisien dan menghasilkan dasar statistik yang kuat untuk proses klasifikasi.

$$TF(t,d) = \frac{(Number\ of\ occurrences\ of\ term\ t\ in\ document\ d)}{(Total\ number\ of\ terms\ in\ the\ document\ d)}$$

$$IDF(t,D) = \log_e \frac{(Total\ number\ of\ documents\ in\ the\ corpus)}{(Number\ of\ documents\ with\ term\ t\ in\ them)}$$

$$TF\text{-}IDF(t,d,D) = TF(t,d) \times IDF(t,D)$$

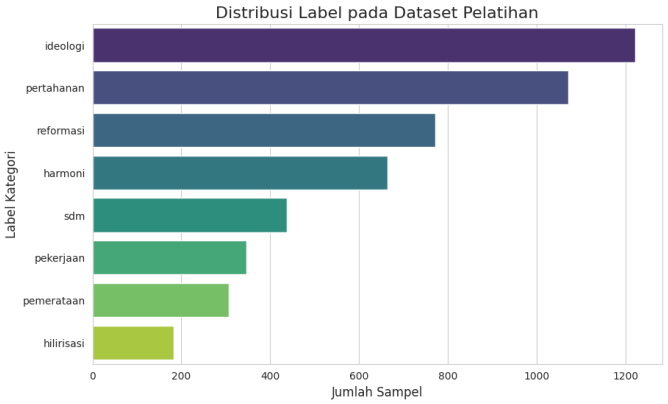
Gambar 2. Mekanisme TF-IDF [1]

D. Eksplorasi dan Visualisasi Data

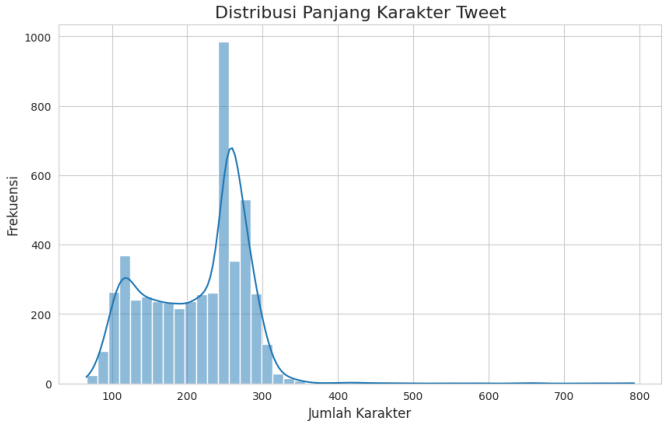
Tahap awal dalam penelitian ini adalah melakukan analisis data eksplorasi (EDA) untuk memahami karakteristik dan tantangan yang melekat pada dataset. Analisis ini menjadi fondasi untuk menentukan strategi pemodelan yang paling tepat.

1) Distribusi Label: Visualisasi distribusi label pada data latih disajikan pada Gambar 3. Dari grafik tersebut, terlihat jelas adanya ketidakseimbangan kelas (*class imbalance*) yang signifikan. Kategori seperti `ideologi` dan `pertahanan` memiliki jumlah sampel yang jauh lebih banyak dibandingkan kategori minoritas seperti `pemerataan` dan `hilirisasi`. Observasi ini sangat krusial dan menjadi justifikasi utama mengapa metrik evaluasi *Balanced Accuracy* digunakan, serta mendorong penerapan teknik *Stratified K-Fold* untuk memastikan proses validasi yang adil dan representatif bagi semua kelas.

2) Analisis Panjang Teks: Distribusi panjang karakter per tweet dianalisis untuk memahami sifat data teks yang akan diolah. Seperti ditunjukkan pada Gambar 4, sebagian besar tweet berada di bawah 280 karakter, yang merupakan karakteristik khas dari platform X. Ini mengkonfirmasi bahwa model harus mampu menangani teks pendek (*short text*) yang memiliki konteks terbatas. Berdasarkan distribusi ini, panjang input maksimal untuk *tokenizer* ditetapkan sebesar 128 token, sebuah nilai yang cukup untuk mencakup mayoritas data tanpa kehilangan informasi signifikan.



Gambar 3. Distribusi Sampel per Kategori Label



Gambar 4. Distribusi Panjang Karakter pada Tweet

3) Analisis Kata Kunci per Kategori: Untuk mendapatkan pemahaman kualitatif tentang setiap topik, dilakukan analisis *n-gram* guna mengidentifikasi frasa yang paling sering muncul untuk setiap kategori. Ditemukan bahwa banyak kata kunci umum seperti ”digital” dan ”pemerintah” muncul di beberapa kategori. Fenomena tumpang tindih kata kunci ini mengindikasikan bahwa model klasifikasi sederhana yang hanya berbasis frekuensi kata (seperti TF-IDF) mungkin akan mengalami kesulitan. Hal ini memperkuat keputusan untuk menggunakan model Transformer seperti `IndoBERTweet` yang mampu memahami konteks kalimat secara mendalam, tidak hanya berdasarkan kata kunci individual. Untuk melihat kata kunci yang sering muncul pada teks, dapat dilihat pada Tabel I.

TABLE I
TOP 5 BIGRAM (KATA KUNCI) PER KATEGORI

Kategori	Top 5 Bigram (Kata Kunci)
Harmoni	umat islam toleransi beragama lingkungan hidup non muslim umat beragama
Hilirisasi	bahan mentah sumber daya ekspor bahan daya alam lapangan kerja
Ideologi	kebebasan berpendapat hak asasi asasi manusia keadilan sosial kebebasan berekspresi
Pekerjaan	ekonomi kreatif lapangan kerja keadilan sosial ekonomi digital rantai pasok
Pemerataan	keadilan sosial akses internet sosial seluruh seluruh rakyat rakyat indonesia
Pertahanan	energi bersih pertahanan negara masa depan keamanan siber energi terbarukan
Reformasi	data pribadi keadilan sosial perlindungan data hak asasi asasi manusia
SDM	akses internet judi online keadilan sosial salah satu bijak bermedia

E. Pemodelan IndoBERT

IndoBERT dipilih sebagai model utama karena telah terbukti unggul dalam memahami bahasa Indonesia melalui representasi kontekstual berbasis transformer [9]. Model ini mampu menangkap hubungan antar kata dalam kalimat, bahkan pada struktur sintaksis yang kompleks sehingga relevan untuk tugas klasifikasi teks berbasis bahasa alami [4] [11].

Pada penelitian ini, IndoBERT tidak digunakan secara mentah, tetapi dilakukan proses *fine-tuning* dengan menyesuaikan bobot model terhadap data latih yang tersedia. Agar pelatihan lebih stabil, panjang *input* teks dibatasi hingga 128 token, yang merupakan kompromi antara efisiensi komputasi dan kemampuan model menangkap konteks. Proses pelatihan dilakukan menggunakan skema validasi silang (5-fold cross-validation) sehingga setiap *subset* data mendapat kesempatan sebagai data validasi. Pendekatan ini memastikan model tidak hanya unggul pada data latih tertentu, tetapi juga memiliki kemampuan generalisasi yang baik terhadap data uji.

Hasil akhir dari tahap ini adalah model IndoBERT *Tuned*, yakni IndoBERT yang telah disesuaikan dengan data spesifik penelitian. Model ini berfungsi sebagai *baseline* model yang menjadi dasar perbandingan terhadap pendekatan lain, sekaligus fondasi untuk mengintegrasikan teknik lanjutan seperti *Confidence Learning* dan *Pseudo-Labelling* pada tahap berikutnya.

F. Confidence Learning

Confidence Learning digunakan sebagai strategi untuk meningkatkan kualitas hasil klasifikasi model dasar (IndoBERT *Tuned*). Pendekatan ini berfokus pada pengelolaan prediksi model berdasarkan tingkat keyakinan (*confidence score*) yang dihasilkan. Alih-alih menerima semua prediksi secara langsung, sistem memberikan bobot atau perlakuan berbeda pada sampel dengan tingkat keyakinan tinggi dan rendah [5].

Secara teknis, model menghasilkan distribusi probabilitas untuk setiap kelas melalui fungsi softmax. Nilai probabilitas ini kemudian dievaluasi: prediksi dengan skor di atas ambang batas tertentu dianggap meyakinkan, sedangkan prediksi di bawah ambang batas dapat diperlakukan sebagai data berisiko tinggi. Proses ini membantu mengurangi propagasi kesalahan yang mungkin terjadi akibat prediksi yang terlalu spekulatif.

Keunggulan utama dari *Confidence Learning* adalah peningkatan reliabilitas model. Dengan memilah prediksi berdasarkan tingkat kepastian, model tidak hanya mengutamakan akurasi tetapi juga kepercayaan terhadap hasil klasifikasi. Hal ini sangat penting pada konteks penelitian ini, mengingat data teks sering kali mengandung ambiguitas dan variasi bahasa yang luas.

Pada akhirnya, *Confidence Learning* berfungsi sebagai lapisan tambahan di atas model IndoBERT, yang tidak hanya menilai benar atau salahnya sebuah prediksi, tetapi juga mengukur seberapa besar keyakinan model terhadap keputusan tersebut. Strategi ini diharapkan dapat menghasilkan model yang lebih *robust* dan dapat diandalkan pada aplikasi nyata. Berikut adalah contoh data tweet yang diidentifikasi memiliki potensi salah oleh *cleanlab*

G. Pseudo-Labelling

Pseudo-Labelling diterapkan sebagai strategi *semi-supervised learning* untuk memperluas jumlah data latih tanpa harus melakukan anotasi manual tambahan [10]. Pada tahap ini, model IndoBERT yang telah melalui proses *fine-tuning* digunakan untuk memprediksi label pada himpunan data uji atau data tidak berlabel.

Hanya prediksi dengan tingkat probabilitas di atas ambang batas tertentu yang dipertahankan dan ditambahkan kembali ke himpunan data latih sebagai *pseudo-label*. Prediksi dengan tingkat keyakinan rendah diabaikan agar tidak menurunkan kualitas data. Dengan demikian, jumlah sampel latih meningkat secara signifikan, tetapi tetap terjaga akurasinya.

Pendekatan ini memberikan dua keuntungan utama. Pertama, model dapat memanfaatkan informasi dari data yang sebelumnya tidak digunakan sehingga memperkaya distribusi kelas dalam data latih. Kedua, strategi ini mendorong model untuk beradaptasi lebih baik terhadap variasi data yang ada di dunia nyata.

Dengan diterapkannya *pseudo-labelling*, dihasilkan model yang tidak hanya bergantung pada data latih berlabel, tetapi juga mampu mengintegrasikan data tambahan secara adaptif. Hal ini diharapkan memperkuat performa model pada tahap evaluasi akhir.

TABLE II
CONTOH TWEET YANG DIIDENTIFIKASI MEMILIKI POTENSI SALAH LABEL OLEH CLEANLAB

ID	Teks Tweet (Disingkat)	Label Asli	Prediksi OOF
4148	*"...keadilan toleransi... keadilan sosial toleransi dan hidup damai..."*	ideologi	**harmoni**
813	*"Jika Tiongkok menginvasi Taiwan... Perang regional mungkin terjadi..."*	hilirisasi	**pertahanan**
1451	*"...selalu ngajarin toleransi pdahal kalo gasuka yaa gapapa itu hak..."*	pertahanan	**harmoni**
2803	*"...transisi energi... berorientasi pada ketahanan energi nasional..."*	hilirisasi	**pertahanan**
4388	*"...debat dosa itu ya kata AI gaada pelanggaran HAM... muterbalik fakta..."*	reformasi	**ideologi**
1645	*"...lebih ngerti soal energi bersih berkat informasi dari pemerintah."*	harmoni	**pertahanan**
785	*"Ini yg namanya negara? Keadilan sosial? Paling pancasila?..."*	hilirisasi	**ideologi**
1721	*"perang siber dan perubahan iklim mempengaruhi keamanan internasional..."*	reformasi	**pertahanan**

H. Penanganan Kelas Tidak Seimbang Dengan Focal Loss

Salah satu tantangan utama dalam penelitian ini adalah distribusi label yang tidak seimbang, di mana kelas mayoritas mendominasi jumlah sampel. Ketidakseimbangan ini berpotensi menyebabkan model bias terhadap kelas mayoritas sehingga performa klasifikasi pada kelas minoritas menurun secara signifikan.

Untuk mengatasi hal tersebut, penelitian ini mengadopsi fungsi kerugian *focal loss*. Berbeda dengan *cross-entropy loss standar*, *focal loss* menambahkan faktor penyesuaian yang memperbesar kontribusi sampel dengan tingkat kesalahan tinggi (*hard-to-classify*) dan mengurangi pengaruh sampel dengan tingkat kepercayaan tinggi (*easy-to-classify*) [3] [13] [7]. Dengan demikian, model dapat lebih fokus belajar dari data minoritas tanpa sepenuhnya mengabaikan kelas mayoritas.

Formulasi matematis dari *focal loss* diberikan sebagai berikut:

$$FL(p_t) = -\alpha_t(1-p_t)^{\gamma}\log(p_t)$$

, di mana p_t adalah probabilitas prediksi benar, α_t merupakan faktor penyeimbang antar kelas, dan γ adalah parameter yang mengatur tingkat penekanan pada sampel sulit.

Implementasi *focal loss* dalam penelitian ini diterapkan terutama pada model *Logistic Regression* berbasis TF-IDF, dengan tujuan meningkatkan sensitivitas model terhadap kelas minoritas. Dengan strategi ini, performa klasifikasi diharapkan

lebih merata pada seluruh label target, bukan hanya terfokus pada kelas dominan.

I. Pra-pemrosesan Data

Sebelum data teks dapat digunakan untuk pemodelan, perlu dilakukan serangkaian tahap pra-pemrosesan. Tujuannya adalah untuk membersihkan teks dari *noise* (informasi yang tidak relevan) dan menstandarisasi formatnya agar dapat diolah secara efektif oleh model. Tahapan yang dilakukan meliputi *case folding* (mengubah semua huruf menjadi huruf kecil), penghapusan URL, eliminasi *mention*, *hashtag*, dan angka, serta pembersihan karakter spesial. Proses ini diilustrasikan pada Tabel III menggunakan salah satu contoh data.

TABLE III
CONTOH TAHAPAN PRA-PEMROSESAN TEKS

Proses	**Status**	**Teks**
Case Folding	Sebelum	`@CM_Community19 @bitgetglobal Smarter Eyes Challenge bukan cuma permainan tapi simulasi nyata dari ancaman digital yang mengintai`
	Sesudah	`@cm_community19 @bitgetglobal smarter eyes challenge bukan cuma permainan tapi simulasi nyata dari ancaman digital yang mengintai`
No URLs	Sebelum	`@cm_community19 @bitgetglobal smarter eyes challenge...`
	Sesudah	`@cm_community19 @bitgetglobal smarter eyes challenge...`
No mentions/ hashtags/ numbers	Sebelum	`@cm_community19 @bitgetglobal smarter eyes challenge...`
	Sesudah	`smarter eyes challenge bukan cuma permainan tapi simulasi nyata dari ancaman digital yang mengintai`
No special chars	Sebelum	`smarter eyes challenge bukan cuma permainan...`
	Sesudah	`smarter eyes challenge bukan cuma permainan tapi simulasi nyata dari ancaman digital yang mengintai`

Setelah melalui tahapan pra-pemrosesan ini, data teks yang bersih siap untuk ditransformasi menjadi representasi numerik untuk digunakan dalam pemodelan.

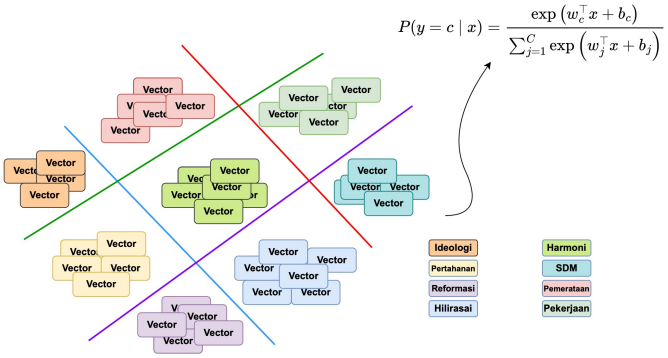
J. Pemodelan Logistic Regression

Selain model berbasis transformer, penelitian ini juga mengimplementasikan *Logistic Regression* sebagai pendekatan pembanding sekaligus komplementer. Model ini dipilih karena kesederhanaannya dalam menangani representasi vektor teks serta kemampuannya memberikan interpretasi berbasis bobot fitur [12].

Dalam tahap feature engineering, teks diubah menjadi representasi numerik menggunakan metode TF-IDF. Strategi ini memungkinkan penekanan kata-kata yang jarang muncul, tetapi memiliki daya diskriminasi tinggi terhadap label target.

Untuk mengatasi masalah ketidakseimbangan kelas, digunakan teknik focal loss yang menitikberatkan pembelajaran pada sampel minoritas. Dengan demikian, model tidak hanya mendominasi kelas mayoritas, tetapi juga lebih sensitif terhadap kelas yang jarang muncul.

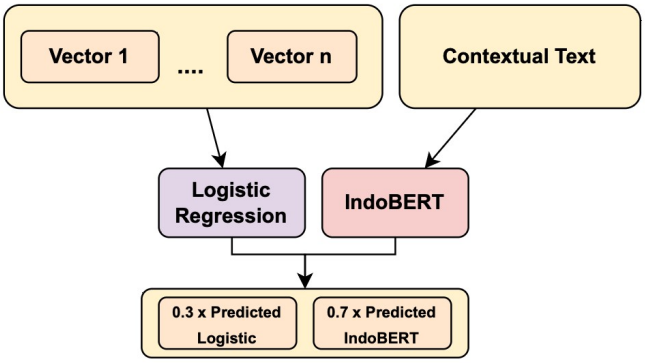
Pelatihan dilakukan menggunakan *stratified 5-fold cross-validation* guna menjaga proporsi distribusi kelas di setiap lipatan sekaligus mengurangi risiko *overfitting*. Hasil akhir model *Logistic Regression* kemudian disiapkan untuk digabungkan dalam strategi *hybrid ensemble* Alur kerja dari strategi ini diilustrasikan pada Gambar 5.



Gambar 5. Diagram Alur Kerja Strategi *Logistic Regression*

K. Strategi Pemodelan Ensemble

Untuk meningkatkan kinerja klasifikasi, penelitian ini menggabungkan dua pendekatan berbeda melalui strategi *hybrid ensemble*. Ide utama dari *ensemble* adalah memanfaatkan keunggulan masing-masing model agar menghasilkan prediksi yang lebih robust dan akurat. Alur kerja dari strategi ini diilustrasikan pada Gambar 6.



Gambar 6. Diagram Alur Kerja Strategi *Hybrid Ensemble*

Seperti yang ditunjukkan pada diagram, *pipeline* ini terdiri dari dua cabang pemodelan utama yang berjalan secara paralel.

Cabang pertama adalah model IndoBERTweet, yang memiliki keunggulan dalam memahami konteks semantik secara mendalam berkat arsitektur berbasis transformer. Model ini dilatih menggunakan metode *Stratified 5-Fold Cross-Validation* untuk meningkatkan stabilitas dan generalisasi, di mana

prediksi akhir dari kelima model tersebut digabungkan melalui *soft voting*.

Cabang kedua adalah model klasik *Logistic Regression*, yang lebih sederhana namun mampu memanfaatkan representasi statistik kata secara efektif melalui TF-IDF. Model ini berfungsi sebagai ”penasihat kata kunci” yang memberikan perspektif berbeda dari model Transformer.

Metode penggabungan akhir yang digunakan adalah *weighted averaging* (blending), di mana probabilitas prediksi dari kedua cabang digabungkan. Bobot prediksi ditetapkan sebesar 0.7 untuk *ensemble* IndoBERTweet dan 0.3 untuk Logistic Regression. Pembobotan ini didasarkan pada performa relatif kedua pendekatan pada tahap validasi, dengan memberikan porsi lebih besar kepada IndoBERTweet sebagai model utama.

Hasil akhir dari proses ini berupa model *hybrid ensemble* yang memanfaatkan kekuatan komplementer kedua pendekatan, dan digunakan untuk menghasilkan prediksi final pada data uji.

III. Analisis dan Pembahasan

Pada bab ini, akan disajikan hasil dari setiap tahap eksperimen yang telah dilakukan. Analisis difokuskan pada perbandingan performa antar metode untuk memberikan justifikasi atas setiap keputusan dalam alur kerja, serta interpretasi terhadap temuan yang diperoleh untuk menjawab tujuan penelitian.

A. Evaluasi dan Perbandingan Performa Model

Untuk menemukan pendekatan terbaik, serangkaian eksperimen dilakukan secara iteratif. Setiap tahap dibangun di atas temuan dari tahap sebelumnya, dengan tujuan untuk secara bertahap meningkatkan skor *Balanced Accuracy*. Hasil dari setiap eksperimen dirangkum dalam Tabel IV.

B. Dari Baseline ke Arsitektur Transformer

Eksperimen dimulai dengan model statistik `TF-IDF + Logistic Regression` yang menghasilkan skor *baseline* sebesar 0.503. Selanjutnya, dilakukan peralihan ke arsitektur Transformer dengan *fine-tuning* model `IndoBERTweet`. Langkah ini menghasilkan lompatan performa yang drastis ke 0.577, yang membuktikan bahwa kemampuan model dalam memahami konteks kalimat sangat krusial untuk tugas klasifikasi teks media sosial.

C. Optimasi Awal dan Hipotesis Kualitas Data

Langkah selanjutnya berfokus pada optimasi dan augmentasi data. Penerapan *pseudo-labeling* sederhana, *hyperparameter tuning*, dan *Stratified K-Fold* secara bertahap berhasil mendorong performa hingga mencapai skor 0.592. Namun, peningkatan skor mulai melandai. Setelah mencoba berbagai model lain tanpa hasil yang signifikan, muncul hipotesis bahwa kualitas label pada data latih menjadi faktor penghambat utama.

TABLE IV
Evolusi Performa Model Melalui Eksperimen Prosedural

No.	Tahap Eksperimen	Deskripsi Metode	Skor
1.	Baseline Awal	`TF-IDF + LogReg` dengan pra-pemrosesan dasar.	0.503
2.	Penggantian Arsitektur	*Fine-tuning* `IndoBERTweet` (4 epoch) sebagai model dasar baru.	0.577
3.	Augmentasi Awal	Menambahkan *pseudo-label* sederhana pada `IndoBERTweet` dasar.	0.585
4.	Optimasi & Augmentasi Lanjutan	Menerapkan HPT, lalu *pseudo-labeling* dengan model hasil *tuning*.	0.590
5.	Validasi Silang	Menerapkan *Stratified K-Fold* (CV=5) pada model hasil HPT.	0.592
6.	**Pembersihan Data**	Melatih ulang K-Fold pada data latih yang sudah dibersihkan oleh `cleanlab`.	**0.595**
7.	Ensembling Hibrida Awal	Menggabungkan (stacking) hasil Langkah 6 dengan `TF-IDF + LogReg`.	0.602
8.	Augmentasi pada Data Bersih	K-Fold pada data bersih yang ditambah *pseudo-label* baru.	0.606
9.	Ensembling Hibrida Final	Menggabungkan (blending) hasil Langkah 8 dengan `TF-IDF + LogReg`.	0.609
10.	**Optimasi Loss**	Menambahkan **Focal Loss** pada *pipeline* Langkah 9.	**0.613**

D. Analisis Kunci: Penanganan Noisy Label dengan Confident Learning

Untuk membuktikan hipotesis tersebut, diterapkan pendekatan *Confident Learning*. `cleanlab` berhasil mengidentifikasi dan menghapus data yang kemungkinan salah label. Pelatihan ulang *pipeline* K-Fold pada dataset yang sudah dibersihkan ini (Langkah 6) berhasil memberikan peningkatan skor menjadi 0.595. Ini secara kuantitatif memvalidasi bahwa pembersihan data dari *noise* label merupakan langkah yang sangat efektif.

E. Sinergi Strategi: Augmentasi, Ensembling, dan Focal Loss

Dengan fondasi data yang lebih bersih, strategi-strategi sebelumnya kembali dieksplorasi dan menunjukkan hasil yang jauh lebih baik.

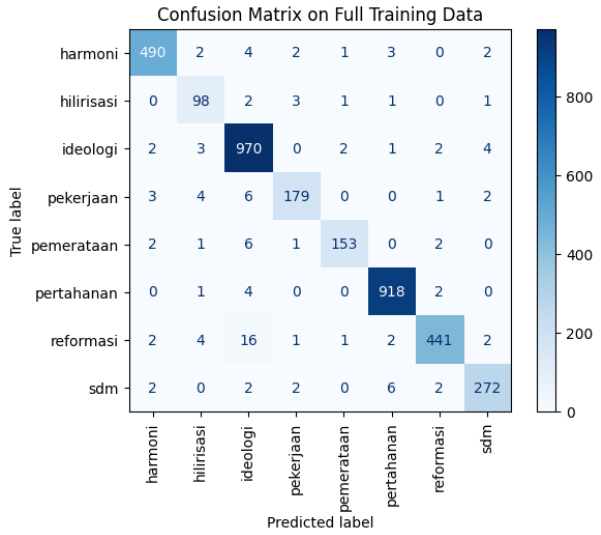
1) Ensembling Hibrida Awal: Menggabungkan prediksi dari *ensemble* K-Fold (pada data bersih) dengan

`TF-IDF + Logistic Regression` berhasil menaikkan skor ke 0.602, membuktikan bahwa model BERT tetap mendapat manfaat dari sinyal kata kunci eksplisit.

2) Augmentasi Lanjutan: *Pseudo-labeling* menggunakan ”guru” yang dilatih pada data bersih berhasil menaikkan skor secara signifikan ke 0.606.
3) Kombinasi Puncak: Menggabungkan augmentasi data bersih dengan *ensembling* hibrida kembali meningkatkan skor ke 0.609.
4) Optimasi *Loss Function*: Sebagai langkah final, *loss function* standar diganti dengan Focal Loss yang dikombinasikan dengan *balanced weights*. Pendekatan ini memaksa model untuk fokus pada contoh-contoh yang sulit dari kelas minoritas, dan berhasil mencapai skor puncak 0.613, yang menjadi hasil akhir dari penelitian ini.

F. Analisis Performa Per-Kelas

Untuk memahami kekuatan dan kelemahan model secara lebih mendalam, dilakukan analisis performa pada setiap kelas menggunakan *Confusion Matrix* dan *Classification Report* dari prediksi *out-of-fold*.



Gambar 7. Confusion Matrix dari Prediksi Out-of-Fold

Confusion Matrix pada Gambar 7 menunjukkan bahwa model memiliki performa yang sangat baik pada kelas-kelas mayoritas seperti `pertahanan`, `ideologi`, dan `harmoni`, yang ditunjukkan oleh tingginya angka pada diagonal utama. Namun, matriks ini juga mengungkap di mana model masih mengalami ”kebingungan”. Terlihat adanya beberapa kesalahan klasifikasi antara kelas yang secara semantik tumpang tindih, seperti sejumlah kecil sampel `pekerjaan` yang salah diklasifikasikan sebagai `sdm`, dan sebaliknya.

Untuk analisis kuantitatif yang lebih detail, Tabel V menyajikan *Classification Report*.

Laporan ini menegaskan bahwa model mencapai performa yang sangat unggul secara konsisten, dengan nilai *precision* dan *recall* melampaui 0.90 pada hampir seluruh kelas. Bahkan pada kelas yang paling menantang, yakni `hilirisasi`, *precision* masih berada pada tingkat tinggi (0.87), sehingga sebagian besar prediksi tetap akurat. Sementara itu, capaian *recall* di atas 0.90, termasuk pada kelas minoritas, menunjukkan kemampuan model yang luar biasa dalam mengenali mayoritas anggota dari setiap kelas dengan benar. Pencapaian ini membuktikan efektivitas penerapan metrik *Balanced Accuracy* serta strategi penanganan ketidakseimbangan data, sehingga keseluruhan model dapat dikategorikan sangat andal dan siap untuk diaplikasikan dalam konteks nyata.

TABLE V
CLASSIFICATION REPORT DARI PREDIKSI OUT-OF-FOLD

	precision	**recall**	**f1-score**	**support**
harmoni	0.98	0.97	0.98	504
hilirisasi	0.87	0.92	0.89	106
ideologi	0.96	0.99	0.97	984
pekerjaan	0.95	0.92	0.93	195
pemerataan	0.97	0.93	0.95	165
pertahanan	0.99	0.99	0.99	925
reformasi	0.98	0.94	0.96	469
sdm	0.96	0.95	0.96	286
accuracy			0.97	3634
macro avg	0.96	0.95	0.95	3634
weighted avg	0.97	0.97	0.97	3634

IV. KESIMPULAN

Penelitian ini berhasil menyusun sebuah *pipeline machine learning end-to-end* yang kuat untuk klasifikasi topik wacana publik Astra Cita pada media sosial X. Melalui pendekatan data-sentris, kami berhasil mengatasi tantangan utama, yaitu data yang tidak seimbang dan label yang *noisy*. Penggunaan model transformer, IndoBERTweet, memberikan lompatan performa yang signifikan, menunjukkan bahwa pemahaman konteks semantik sangat penting. Hipotesis kami bahwa data latih mengandung label yang tidak akurat divalidasi dengan keberhasilan penerapan *Confident Learning*, yang menjadi fondasi untuk langkah-langkah selanjutnya. Kombinasi dari berbagai teknik, seperti *pseudo-labeling* yang diterapkan pada data yang sudah bersih dan *hybrid ensembling* antara IndoBERTweet dan *Logistic Regression*, menunjukkan sinergi yang efektif. Terakhir, penggunaan Focal Loss sebagai fungsi kerugian terbukti krusial dalam menyeimbangkan performa model pada kelas-kelas minoritas. Hasil akhir berupa skor Balanced Accuracy sebesar 0,613 membuktikan bahwa strategi berlapis yang berfokus pada kualitas data dan *ensembling hibrida* lebih unggul daripada sekadar optimasi model tunggal, membuka jalan untuk analisis wacana publik yang lebih akurat dan andal.

Untuk penelitian selanjutnya, akurasi klasifikasi dapat lebih ditingkatkan dengan mengeksplorasi teknik-teknik yang lebih canggih. Salah satu area yang potensial adalah data augmentation generatif, menggunakan model bahasa besar untuk menciptakan sampel-sampel baru yang relevan, terutama untuk kelas minoritas seperti ’hilirisasi’. Selain itu, pengujian model-model Transformer yang lebih besar atau yang secara spesifik dilatih untuk bahasa Indonesia dan teks media sosial dapat

dilakukan untuk melihat apakah ada batas atas performa yang dapat dicapai. Terakhir, pendekatan *ensemble learning* yang lebih kompleks, seperti stacking atau blending dengan bobot yang dioptimalkan, dapat dipertimbangkan untuk meningkatkan akurasi prediksi, terutama pada kelas-kelas yang secara semantik saling tumpang tindih.

REFERENCES

[1] A. H. J. Almarashy, M. R. Feizi-Derakhshi, and P. Salehpour, "Enhancing fake news detection by multi-feature classification," IEEE Access, vol. 11, pp. 139601–139613, 2023.

[2] Asosiasi Penyelenggara Jasa Internet Indonesia (APJII), "Hasil Survei Penetrasi Internet Indonesia 2024," *apjii.or.id*, Jan. 31, 2024. [Online]. Available: https://apjii.or.id/berita/d/apjii-jumlah-pengguna-internet-indonesia-tembus-221-juta-orang. [Accessed: Sep. 1, 2025].

[3] D. Sarkar, A. Narang, and S. Rai, "Fed-focal loss for imbalanced data classification in federated learning," arXiv preprint arXiv:2011.06283, 2020.

[4] F. Muftie and M. Haris, "IndoBERT Based Data Augmentation for Indonesian Text Classification," 2023 International Conference on Information Technology Research and Innovation (ICITRI), Jakarta, Indonesia, 2023, pp. 128-132, doi: 10.1109/ICITRI59340.2023.10250061.

[5] H. Chen, et al., "SoftMatch: Addressing the quantity-quality trade-off in semi-supervised learning," arXiv preprint arXiv:2301.10921, 2023.

[6] IBM, "Apa itu NLP (Natural Language Processing atau Pemrosesan Bahasa Alami)?," *ibm.com*, Aug. 2, 2025. [Online]. Available: https://www.ibm.com/id-id/think/topics/natural-language-processing. [Accessed: Sep. 1, 2025].

[7] J. Dong, "Focal loss improves the model performance on multi-label image classifications with imbalanced data," in Proc. 2nd Int. Conf. Ind. Control Netw. Syst. Eng. Res. (ICNSER), New York, NY, USA, 2020, pp. 18–21, doi: 10.1145/3411016.3411020.

[8] Kementerian PPN/Bappenas, "Indonesia Emas 2045," *indonesia2045.go.id*, 2024. [Online]. Available: https://indonesia2045.go.id/. [Accessed: Sep. 1, 2025].

[9] Koto, F., Rahimi, A., Lau, J. H., Baldwin, T. (2020). IndoLEM and IndoBERT: A benchmark dataset and pre-trained language model for Indonesian NLP. arXiv preprint arXiv:2011.00677.

[10] M. Ahmed, B. Wen, A. Luo, S. Pan, J. Su, X. Cao, and Y. Liu, "Towards robust learning with noisy and pseudo labels for text classification," Information Sciences, vol. 661, p. 120160, 2024, doi: 10.1016/j.ins.2024.120160.

[11] M. R. Syazali and E. Yulianti, "Classification of economic activities in Indonesia using IndoBERT language model," Jurnal Ilmu Komputer dan Informasi, vol. 18, no. 2, pp. 155–165, 2025.

[12] T. Pranckevičius and V. Marcinkevičius, "Comparison of naive bayes, random forest, decision tree, support vector machines, and logistic regression classifiers for text reviews classification," Baltic Journal of Modern Computing, vol. 5, no. 2, pp. 221–232, 2017.

[13] W. Lin, P. Wu and X. Xiao, "Boundary Focal Loss for Class Imbalanced Learning," 2021 14th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), Shanghai, China, 2021, pp. 1-5, doi: 10.1109/CISP-BMEI53629.2021.9624339.