applied sciences

MDPI

Article

A Hybrid Feature-Enhanced IndoBERT Framework with Controlled Semi-Supervised Learning for Low-Resource Indonesian Hate Speech Detection

Shoffan Saifullah [1,2,*] and Rafał Dreżewski [1,*]

[1] Faculty of Computer Science, AGH University of Krakow, 30-059 Krakow, Poland
[2] Department of Informatics, Universitas Pembangunan Nasional Veteran Yogyakarta, Yogyakarta 55281, Indonesia
* Correspondence: saifulla@agh.edu.pl or shoffans@upnyk.ac.id (S.S.); drezew@agh.edu.pl (R.D.)

Abstract

Low-resource hate speech detection remains a challenging task for Indonesian social media due to limited labeled annotations, highly informal linguistic expressions, and substantial lexical variability. Under such conditions, purely supervised transformer models often suffer from unstable semantic generalization, while conventional pseudo-labeling methods are vulnerable to noisy unlabeled sample propagation. To address these limitations, this study proposes a hybrid feature-enhanced IndoBERT framework integrated with a controlled semi-supervised learning strategy. The proposed model combines contextual IndoBERT embeddings with abusive lexicon cues, handcrafted linguistic indicators, and TF-IDF–SVD statistical representations through a lightweight concatenation–projection feature fusion mechanism, while unlabeled data are incorporated via adaptive confidence thresholding and class-balanced pseudo-label selection to improve pseudo-label reliability. Extensive experiments were conducted under realistic low-resource supervision settings using only 5%, 10%, and 20% labeled data, and the proposed framework was systematically compared against representative baselines, including sparse lexical machine learning models, shallow neural architectures, multilingual transformers, IndoBERTweet, naive pseudo-labeling, and LLM-based prompting. The results show that model effectiveness is strongly supervision-dependent. Under the most extreme low-resource setting, compact statistical augmentation provides the most stable complementary signal, whereas under moderate low-resource supervision, the full hybrid representation combined with controlled semi-supervised learning yields the strongest and most consistent gains. The proposed Hybrid IndoBERT + controlled SSL framework outperforms all baselines at the 20% labeled setting, reaching an accuracy of 0.8654, Macro-F1 of 0.8633, and ROC-AUC of 0.9334. Additional analyses of pseudo-label reliability, calibration behavior, computational efficiency, and qualitative error patterns further show that the proposed framework improves low-resource robustness while maintaining comparable inference-time efficiency. These findings demonstrate that low-resource hate speech detection benefits most from the staged integration of contextual semantic modeling, interpretable linguistic cues, global lexical–statistical structure, and carefully regulated unlabeled data exploitation. Additional experiments using GPT-4o-mini and Llama-3.1-8B further demonstrate that the proposed framework remains competitive against general-purpose large language model prompting approaches under low-resource Indonesian hate speech detection scenarios. The proposed framework provides a practical and reproducible direction for hate speech detection in annotation-constrained social media environments.

Check for updates

Academic Editor: Pedro Couto

Received: 6 May 2026
Revised: 14 June 2026
Accepted: 23 June 2026
Published: 29 June 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

Keywords: hate speech detection; low-resource NLP; IndoBERT; hybrid feature fusion; semi-supervised learning; pseudo-labeling; TF-IDF; Indonesian language processing

1. Introduction

The rapid growth of social media platforms has significantly increased the spread of user-generated content, including harmful and offensive language such as hate speech [1,2]. In the Indonesian context, hate speech detection presents unique challenges due to the informal nature of online communication, frequent use of slang, code-mixing, and the limited availability of high-quality annotated datasets [3–5]. These characteristics make the task particularly difficult for conventional natural language processing (NLP) approaches, especially under low-resource conditions where labeled data is scarce [6,7]. In practice, the scarcity of annotated hate speech corpora often forces classifiers to operate with incomplete semantic coverage, weak minority-class representation, and unstable lexical generalization, thereby reducing the reliability of automatic moderation systems.

Recent advances in deep learning, particularly transformer-based models such as BERT and its variants, have demonstrated strong performance in various NLP tasks by capturing contextual semantic representations [8–10]. IndoBERT, a pre-trained language model specifically designed for the Indonesian language, has shown promising results in text classification tasks [11,12]. However, despite their effectiveness, transformer-based models still rely heavily on sufficiently large labeled corpora to form robust semantic boundaries and may struggle to generalize effectively in low-resource scenarios [13,14]. Moreover, purely contextual models primarily emphasize latent semantic dependencies while often underutilizing explicit lexical cues, abusive lexicons, punctuation styles, stylistic irregularities, and corpus-level frequency patterns that remain highly informative in hate speech classification [15,16].

To address these limitations, hybrid approaches that integrate deep contextual representations with handcrafted and statistical features have gained increasing attention. Handcrafted linguistic features, including lexical indicators, punctuation patterns, and stylistic signals, can provide complementary information that is not fully captured by transformer embeddings [17–19]. Similarly, traditional sparse representations such as TF-IDF, especially when combined with dimensionality reduction techniques such as Singular Value Decomposition (SVD), can preserve global structural and frequency-based characteristics of text [20,21]. Nevertheless, most existing hybrid hate speech detection models still treat these heterogeneous feature sources as auxiliary additions rather than as supervision-sensitive components whose effectiveness may vary substantially under different levels of labeled data availability.

Another promising direction for improving low-resource performance is semi-supervised learning (SSL), which leverages unlabeled data to enhance model training [22]. Among SSL strategies, pseudo-labeling is one of the most widely adopted approaches, where confident predictions on unlabeled samples are iteratively incorporated as additional training data [7,23]. However, naive pseudo-labeling remains highly vulnerable to confidence instability, class imbalance reinforcement, and noisy label propagation, particularly when the initial supervised model is trained on only a small fraction of annotated data [24–26]. As a result, simply enlarging the training set with pseudo-labels does not necessarily translate into stable generalization gains. Designing a controlled and reliability-aware SSL mechanism therefore remains an open challenge for practical low-resource hate speech detection.

Motivated by these limitations, this study proposes a hybrid feature-enhanced IndoBERT framework for low-resource Indonesian hate speech detection, integrating transformer-based contextual embeddings with handcrafted linguistic features and TF-IDF–SVD statistical representations through a structured fusion architecture. The proposed approach is further extended with a controlled semi-supervised learning strategy that incorporates adaptive confidence thresholding and class-balanced pseudo-label selection in order to regulate pseudo-label quality and suppress error accumulation during iterative training.

Extensive experiments are conducted on a merged dataset comprising multiple Indonesian hate speech corpora under realistic low-resource supervision settings, specifically using only 5%, 10%, and 20% labeled data [22]. In contrast to many prior studies that report limited baseline comparisons, this work systematically evaluates the proposed framework against classical lexical machine learning models, shallow neural architectures, multilingual transformers, domain-adaptive pretrained transformers, and conventional pseudo-labeling baselines under the same experimental protocol. The results reveal not only consistent superiority of the proposed framework but also an important supervision-dependent transition in feature effectiveness, where compact statistical representations provide the strongest gains under extreme scarcity, while hybrid semantic–statistical learning with controlled SSL becomes dominant under moderate low-resource supervision.

In addition to conventional benchmark comparisons, this study also evaluates the proposed framework against recent large language model (LLM)-based classifiers, including GPT-4o-mini and Llama-3.1-8B. This comparison is intended to examine whether domain-specialized hybrid feature learning and controlled semi-supervised optimization remain competitive relative to modern general-purpose foundation models under low-resource Indonesian hate speech detection settings.

The main contributions of this work are as follows:

- We propose a feature-enhanced IndoBERT framework for low-resource Indonesian hate speech detection that integrates contextual semantic embeddings, handcrafted linguistic indicators, and compact TF-IDF–SVD statistical representations. In response to the limitation of conventional hybrid concatenation-based classifiers, the proposed framework's novelty lies not in concatenation alone but in the supervision-aware integration and evaluation of heterogeneous feature groups under different labeled-data regimes.
- We introduce a controlled semi-supervised learning strategy based on adaptive confidence thresholding and class-balanced pseudo-label selection. This mechanism is designed to address the common weaknesses of standard self-training and confidence-based pseudo-labeling, particularly noisy pseudo-label propagation, unreliable early confidence estimation, and majority-class reinforcement during iterative unlabeled-data expansion.
- We provide a systematic low-resource evaluation under 5%, 10%, and 20% labeled-data settings, comparing the proposed framework with representative lexical machine learning models, shallow neural models, Transformer-based classifiers, domain-adaptive Indonesian language models, naive pseudo-labeling baselines, and large language model prompting baselines. This comparative design clarifies the methodological position of the proposed framework relative to conventional lexical-semantic classifiers, Transformer-only models, and standard semi-supervised alternatives.
- We analyze the supervision-dependent contribution of heterogeneous feature groups through feature ablation and TF-IDF–SVD sensitivity experiments, showing that compact statistical representations are particularly useful under extreme label scarcity, while the full hybrid representation combined with controlled semi-supervised learning becomes more effective when moderate labeled supervision is available. This

analysis helps explain why the observed gains should be interpreted in terms of stability, feature complementarity, and low-resource robustness rather than only absolute performance improvement.

- We examine the practical reliability of the proposed framework through additional analyses of pseudo-label acceptance behavior, class-distribution preservation, calibration quality, computational efficiency, and qualitative error cases. These analyses provide a more transparent assessment of pseudo-label reliability, model complexity, and the behavior of the proposed framework on ambiguous, implicit, slang-based, and difficult hate speech expressions.

The remainder of this paper is organized as follows. Section 2 reviews the related literature on hate speech detection, transformer-based models, feature fusion techniques, and semi-supervised learning approaches. Section 3 presents the proposed hybrid feature-enhanced IndoBERT framework and the controlled semi-supervised learning strategy. Section 4 describes the experimental setup, including datasets, preprocessing, and evaluation metrics. Section 5 provides a comprehensive analysis of the experimental results and discusses the performance under low-resource scenarios. Finally, Section 6 concludes the paper and outlines future research directions.

2. Related Work

This section reviews the existing literature on hate speech detection, transformer-based models, hybrid feature fusion strategies, and semi-supervised learning approaches. Particular attention is paid to the limitations of current methods in low-resource scenarios and their implications for Indonesian language processing.

2.1. Hate Speech Detection

Hate speech detection has become an important research topic in natural language processing due to the rapid increase in harmful user-generated content on social media platforms [22,27]. Early computational approaches predominantly relied on conventional machine learning classifiers such as Support Vector Machines (SVMs), Logistic Regression, Naïve Bayes, and Random Forest [28,29], typically using sparse lexical representations including bag-of-words, character n-grams, word n-grams, and TF-IDF features [30–33]. These models are computationally efficient and often provide reasonable interpretability, particularly when offensive expressions are explicitly manifested through recurring lexical patterns [34,35]. Several studies have reported that such feature-based approaches remain competitive for noisy social media corpora because sparse lexical boundaries can be learned even from relatively limited training samples [17,36,37].

However, the hate speech detection problem extends beyond simple keyword spotting [38,39]. Hate expressions are frequently conveyed through implicit hostility, sarcastic framing, metaphorical insults, identity-based insinuations, and context-dependent derogatory references [5]. Under these circumstances, purely lexical models exhibit limited capability in capturing semantic dependency, pragmatic meaning, and discourse-level contextual nuance [40,41]. Their decision boundaries are often dominated by surface word frequency, making them sensitive to vocabulary shift and weak in handling semantically equivalent but lexically diverse abusive expressions.

These challenges become substantially more severe in Indonesian social media environments. Indonesian online communication is characterized by highly informal sentence construction, phonetic spelling variations, repeated character emphasis, slang abbreviations, regional vocabulary, and frequent code-mixing with English or local dialects [42,43]. In addition, offensive discourse is often expressed through indirect formulations, abbreviations, or culturally contextualized references that are difficult to capture using fixed lexical

dictionaries alone. As reported in recent Indonesian hate speech studies [3–5], this linguistic variability introduces substantial representational sparsity and increases the ambiguity between hateful and non-hateful utterances.

Another major limitation in existing hate speech detection research is the dominant assumption of fully supervised training. A large portion of prior studies evaluate models under settings where sufficiently annotated corpora are available, enabling stable supervised optimization [44,45]. However, in practice, hate speech annotation is expensive, subjective, and time-consuming, particularly for Indonesian data, where linguistic nuance and cultural sensitivity require careful human judgment [38,46]. Consequently, many real-world deployments must operate under limited labeled data, making conventional fully supervised optimization unreliable. Under such low-resource settings, sparse lexical models may retain some stability, but they often fail to generalize beyond explicit abusive vocabulary and struggle to capture deeper contextual hostility [14,47,48].

These observations indicate that hate speech detection in Indonesian low-resource scenarios requires a model that is simultaneously capable of: (i) preserving the robustness of frequency-based lexical discrimination, (ii) learning contextual semantic meaning beyond explicit keywords, and (iii) maintaining stable performance when annotated supervision is highly constrained. This need motivates the exploration of more advanced contextual representation models and feature-enhanced learning frameworks beyond conventional machine learning pipelines.

2.2. Transformer-Based Models for Text Classification

The introduction of transformer architectures has substantially changed the landscape of text classification by enabling models to learn contextualized token representations through self-attention mechanisms [49]. Unlike conventional sparse lexical models that treat words as largely independent frequency units, transformer-based models capture bidirectional semantic dependencies, long-range contextual interactions, and subtle compositional meaning within a sentence [50]. This contextual modeling capability has led to significant performance gains in sentiment analysis, abusive language identification, sarcasm detection, and hate speech classification across multiple languages [51].

Among the most influential transformer frameworks, BERT and its derivatives have become standard baselines for supervised text classification [52,53]. These models are pretrained on large unlabeled corpora using masked language modeling objectives, allowing them to encode rich semantic and syntactic priors before downstream fine-tuning. For Indonesian language processing, IndoBERT has shown particularly strong effectiveness because its pretraining corpus reflects local morphological structures, Indonesian word usage patterns, and common semantic compositions more accurately than multilingual models [54,55]. Consequently, IndoBERT frequently outperforms earlier recurrent or convolutional neural architectures in Indonesian sentiment and abusive-text classification tasks [56,57].

Recent studies have further extended this transformer paradigm through multilingual and domain-adaptive pretraining [58,59]. Models such as mBERT and XLM-R provide cross-lingual contextual representations that can transfer knowledge across languages [60], while IndoBERTweet specifically adapts transformer pretraining to Indonesian Twitter-style social media discourse [61]. Such domain-adaptive pretraining is especially relevant for hate speech detection because abusive online communication often differs substantially from formal text corpora in terms of slang, abbreviations, emotive punctuation, and fragmented syntax [56]. Empirical evidence has shown that these specialized transformer models often provide more stable semantic understanding than conventional neural architectures when dealing with noisy social media inputs [62].

Despite these advances, transformer-based hate speech classifiers still exhibit several critical limitations under low-resource supervision [17,63]. First, fine-tuning large contextual encoders remains highly dependent on the quantity and representativeness of labeled data [14,19,64]. When only a small fraction of annotations is available, parameter updates may become unstable, causing overfitting to spurious lexical cues and reducing generalization to unseen expressions. Second, although transformers excel at semantic contextualization, they do not explicitly guarantee the preservation of sparse but highly informative lexical markers, such as abusive identity terms, repeated punctuation aggression, explicit profanity counts, or stylistic hostility patterns [65]. Many of these signals are low-dimensional yet highly discriminative in practical hate speech corpora.

Third, not all pretrained transformers behave equally under Indonesian low-resource conditions. While IndoBERT benefits from language-specific pretraining, multilingual encoders such as mBERT and XLM-R may suffer from representational dilution because Indonesian hate speech characteristics constitute only a small portion of their multilingual pretraining distributions [3,54]. Similarly, even domain-adapted models such as IndoBERTweet remain fundamentally dependent on contextual embedding fine-tuning and do not inherently address the scarcity of supervised labels [56,66]. This indicates that contextual semantic learning alone, although powerful, is insufficient to ensure stable low-resource discrimination.

Therefore, while transformer architectures provide a substantially stronger foundation than conventional lexical classifiers, their practical limitations suggest that additional auxiliary representations are needed to preserve sparse linguistic indicators and stabilize classification under constrained supervision. This motivates the exploration of hybrid feature fusion strategies that combine contextual embeddings with handcrafted and statistical textual descriptors.

2.3. Hybrid Feature Fusion Approaches

To overcome the representational limitations of purely contextual transformer models, recent studies have increasingly explored hybrid learning frameworks that combine deep semantic embeddings with auxiliary handcrafted and statistical features [22,36,41,67]. The central motivation behind this line of work is that hate speech is not solely a semantic phenomenon; it is also frequently characterized by explicit lexical triggers, punctuation aggression, capitalization emphasis, repetitive abusive markers, identity-based slurs, and frequency-driven textual regularities that may not always be optimally preserved within dense contextual embeddings. Consequently, integrating heterogeneous feature sources offers the possibility of constructing a richer and more discriminative textual representation.

Handcrafted linguistic features constitute one major branch of this hybrid paradigm. These include profanity counts, abusive lexicon occurrences, punctuation density, excessive capitalization, elongated character patterns, negation structures, sentiment polarity cues, and stylistic irregularities [17–19]. Such descriptors provide explicit interpretable signals that can directly reflect aggressive or derogatory language behavior. In Indonesian social media text, where abbreviated insults, nonstandard spellings, and emotionally exaggerated punctuation are common, these handcrafted cues may encode information that contextual encoders treat as peripheral or noisy [42].

A second important branch involves sparse statistical text representations such as TF-IDF and n-gram frequency embeddings [20,68]. Although these methods are often considered conventional, they retain a strong ability to model corpus-level discriminative lexical structures. When combined with dimensionality reduction techniques such as Singular Value Decomposition (SVD), TF-IDF vectors can preserve dominant latent semantic-frequency patterns while remaining computationally compact [20,21]. This prop-

erty is particularly useful under low-resource conditions, where stable global frequency information may provide more reliable supervision than heavily parameterized contextual embeddings alone. Several recent studies have shown that such compact statistical descriptors remain surprisingly competitive in hate speech and offensive language classification when annotated data are limited.

Despite this potential, the majority of existing hybrid frameworks employ relatively naive feature concatenation, where contextual embeddings, handcrafted descriptors, and sparse statistical vectors are simply joined into a single representation before classification. This straightforward fusion often leads to several technical problems. First, heterogeneous features naturally differ in scale, density, and statistical distribution. Dense contextual embeddings typically dominate optimization due to their high dimensionality, while low-dimensional handcrafted cues may contribute only marginally unless explicitly normalized and balanced. Second, simple concatenation does not guarantee a meaningful interaction between semantic and lexical feature groups; rather, it frequently produces representation redundancy or underutilization of auxiliary signals. Third, previous studies rarely examine whether the usefulness of each feature group changes as labeled supervision becomes more constrained [69–71].

These limitations indicate that hybrid learning should not be understood merely as feature accumulation but as a structured representation integration problem. An effective hybrid framework must preserve the semantic richness of transformer encoders while simultaneously ensuring that sparse lexical descriptors and handcrafted linguistic cues contribute nontrivially to the final decision boundary. More importantly, the relative importance of these heterogeneous representations may shift under different low-resource regimes, suggesting that feature complementarity itself requires systematic empirical analysis.

Therefore, a carefully designed structured fusion mechanism is necessary to fully exploit contextual, handcrafted, and statistical textual information in a unified classifier. This forms the primary motivation for the feature-enhanced hybrid architecture proposed in this study.

2.4. Advanced Semi-Supervised and Low-Resource NLP Approaches

Beyond feature engineering and contextual representation learning, another major direction for improving low-resource text classification is the utilization of unlabeled data through semi-supervised learning (SSL) [72]. In many practical natural language processing applications, unlabeled textual data is abundant, while expert annotation remains limited and costly. This imbalance has motivated extensive research on SSL strategies such as self-training, pseudo-labeling, consistency regularization, teacher–student learning, and iterative confidence refinement [73]. The central objective of these approaches is to supplement the sparse supervised signals with automatically generated supervisory information derived from the unlabeled corpus.

Among these strategies, pseudo-labeling has become one of the most widely adopted due to its conceptual simplicity and easy integration with existing supervised architectures [74]. In pseudo-labeling, a model trained on a small labeled subset is used to predict labels for unlabeled instances, and highly confident predictions are then recycled as additional training samples [75]. Several modern SSL frameworks, including Noisy Student, FixMatch-inspired confidence filtering, and iterative self-training pipelines, have demonstrated that such automatic label expansion can improve model generalization when the pseudo-label quality is sufficiently reliable [76,77].

However, the practical effectiveness of pseudo-labeling is highly sensitive to the reliability of early model predictions. In severe low-resource settings, the initial supervised classifier is often weak, which causes confidence estimates to be poorly calibrated [78]. As

a result, naive pseudo-labeling can easily introduce mislabeled samples into the training pool, amplifying noise over successive iterations and shifting the decision boundary toward erroneous regions [25]. This phenomenon is especially problematic in hate speech detection, where class boundaries are inherently ambiguous, and many samples contain subtle hostility that even human annotators may interpret inconsistently.

A second major issue concerns class imbalance reinforcement. Standard pseudo-labeling pipelines typically select samples solely based on confidence magnitude without explicitly controlling class distribution [74]. In imbalanced hate speech corpora, this causes the dominant class to contribute disproportionately more pseudo-labeled samples, which gradually biases the classifier toward majority predictions [79]. Consequently, overall accuracy may increase while minority class sensitivity and Macro-F1 remain unstable. Such imbalance amplification has been repeatedly identified as one of the major weaknesses of naive semi-supervised text classification.

Recent low-resource NLP studies have attempted to address these problems through more sophisticated confidence scheduling, uncertainty estimation, or distribution-aware sample filtering [80,81]. Nevertheless, most existing frameworks are designed as general-purpose SSL systems and are not specifically optimized for the interaction between noisy social media language, heterogeneous feature fusion, and binary hate speech imbalance. In particular, few studies explicitly combine pseudo-label reliability control with class-balanced unlabeled sample incorporation within a hybrid transformer-based architecture [82,83].

These observations indicate that unlabeled data cannot simply be injected into low-resource hate speech classifiers through unrestricted pseudo-label expansion. Instead, SSL must be carefully regulated so that additional supervisory signals improve representation learning without causing noise accumulation or majority-class drift. This motivates the development of a controlled semi-supervised learning strategy in which pseudo-label confidence is adaptively filtered and unlabeled sample selection is explicitly balanced across classes before iterative retraining.

2.5. Large Language Models for Low-Resource NLP

Recent large language models (LLMs), including GPT-based and LLaMA-based architectures, have demonstrated strong zero-shot and few-shot capabilities across various NLP tasks [84,85]. These models leverage large-scale pretraining and instruction-following behavior to generalize across multiple domains without task-specific optimization. Nevertheless, their performance in low-resource hate speech detection remains challenging, particularly for culturally contextualized and informal social media language [78,86]. In addition, prompt-based inference may suffer from unstable decision boundaries and reduced sensitivity toward minority linguistic patterns when domain adaptation is absent. These limitations motivate the need for specialized architectures that combine contextual semantic modeling with domain-aware feature representations and controlled semi-supervised learning.

2.6. Research Gap and Motivation

The preceding review reveals that existing hate speech detection approaches still suffer from several unresolved limitations, particularly when deployed under low-resource Indonesian social media conditions. Conventional lexical machine learning models provide relatively stable sparse discrimination but lack contextual semantic understanding, making them weak in handling implicit hostility, sarcasm, and semantically nuanced abusive discourse. Transformer-based architectures substantially improve contextual modeling capability; however, their effectiveness remains highly dependent on sufficient labeled

supervision, and they do not explicitly preserve sparse linguistic aggression markers or corpus-level lexical frequency regularities.

Hybrid feature fusion studies have attempted to bridge this divide by combining contextual embeddings with handcrafted or statistical descriptors. Nevertheless, most prior frameworks rely on simple concatenation-based integration without systematically addressing heterogeneous feature imbalance, representation dominance, or supervision-dependent feature contribution. As a result, the true complementary behavior of contextual, handcrafted, and frequency-based signals under different low-resource regimes remains insufficiently understood.

At the same time, semi-supervised learning offers an attractive avenue for exploiting abundant unlabeled text, yet conventional pseudo-labeling methods are vulnerable to two major issues: unreliable early confidence estimation and progressive class imbalance reinforcement. These weaknesses are particularly damaging for Indonesian hate speech detection, where noisy informal language, ambiguous toxicity boundaries, and skewed class distributions make pseudo-label noise accumulation more likely. Existing SSL frameworks rarely incorporate explicit mechanisms to jointly regulate pseudo-label reliability and balanced unlabeled sample expansion within a hybrid feature-based hate speech classifier.

Therefore, an important methodological gap remains unaddressed in the current literature: there is still no unified low-resource Indonesian hate speech detection framework that simultaneously (i) preserves contextual semantic understanding, (ii) exploits sparse handcrafted linguistic cues, (iii) retains stable global statistical lexical structure, and (iv) utilizes unlabeled data through a controlled pseudo-labeling process that minimizes noise propagation and majority-class drift.

However, the novelty of the present study should not be interpreted as originating from feature concatenation alone, since concatenation-based fusion has already been widely used in hybrid text classification models. Instead, the main methodological contribution lies in the supervision-aware integration of three complementary representation sources within a controlled low-resource learning pipeline. The proposed framework jointly examines contextual semantic information from IndoBERT, explicit linguistic hostility indicators, and compact TF-IDF–SVD statistical descriptors under different labeled-data regimes. In this sense, the framework is designed not only to improve classification performance but also to analyze how the utility of each feature group changes when supervision is severely limited. Furthermore, the semi-supervised component differs from unrestricted pseudo-label expansion by combining adaptive confidence thresholding with class-balanced pseudo-label quotas, thereby reducing two common failure modes in low-resource hate speech detection: noisy pseudo-label propagation and majority-class drift.

Motivated by this gap, the present study introduces a hybrid feature-enhanced IndoBERT framework integrated with a controlled semi-supervised learning strategy. The proposed architecture combines transformer contextual embeddings, handcrafted linguistic descriptors, and TF-IDF–SVD statistical representations through a structured fusion mechanism, while unlabeled data are incorporated via adaptive confidence thresholding and class-balanced pseudo-label selection. Through this unified design, the study aims not only to improve classification performance under limited supervision but also to systematically investigate how heterogeneous feature representations and regulated unlabeled learning interact across different low-resource data regimes.

3. Proposed Method

This section presents the proposed hybrid framework for low-resource Indonesian hate speech detection. The approach integrates deep contextual representations from IndoBERT with complementary handcrafted linguistic features and statistical TF-IDF–SVD

representations to capture both semantic and structural characteristics of textual data. This hybrid design addresses the limitations of purely transformer-based models, which often overlook explicit linguistic cues in noisy and informal social media text.

To further enhance learning under limited labeled data conditions, a controlled semi-supervised learning strategy is introduced. Unlike conventional pseudo-labeling approaches, the proposed method incorporates adaptive thresholding and class-balanced pseudo-label selection to mitigate noise propagation and class imbalance during training. This enables more reliable utilization of unlabeled data while maintaining stable model performance.

The overall framework consists of three main components: (i) a preprocessing and feature extraction module, (ii) a hybrid feature fusion and classification module, and (iii) a controlled semi-supervised learning mechanism for iterative model refinement. These components are described in detail in the following subsections.

3.1. Overview of the Proposed Framework

The proposed method is designed to address two central challenges in low-resource Indonesian hate speech detection: (i) the limited robustness of purely contextual transformer models in preserving explicit lexical hostility indicators, and (ii) the instability of conventional pseudo-labeling when only a small amount of annotated supervision is available. To address these issues, we introduce a unified hybrid framework that combines structured feature-enhanced representation learning with a controlled semi-supervised training mechanism.

Given an input text sample x, the framework first applies textual normalization to obtain a cleaned representation $x' = \mathcal{P}(x)$, which reduces orthographic noise, informal lexical variation, and social-media-specific inconsistencies. The normalized text is then processed through two parallel representational branches. The first branch generates deep contextual semantic embeddings using a pretrained IndoBERT encoder, while the second branch constructs complementary auxiliary numerical descriptors derived from handcrafted linguistic indicators and TF-IDF–SVD statistical representations. These heterogeneous features are subsequently normalized and fused into a unified classification vector.

Unlike conventional single-encoder classifiers that rely exclusively on contextual semantic abstraction, the proposed dual-branch design explicitly preserves both implicit semantic dependencies and sparse lexical aggression markers. This is particularly important for Indonesian hate speech analysis, where offensive intent is often conveyed not only through contextual sentence meaning but also through explicit profanity counts, identity slurs, punctuation emphasis, repeated-character aggression, and frequency-based textual regularities.

To further improve robustness under limited labeled supervision, the framework is extended with a controlled semi-supervised learning loop. Instead of directly adopting naive pseudo-label expansion, unlabeled samples are incorporated only when their prediction confidence satisfies adaptive threshold constraints and when pseudo-label selection remains approximately class balanced. This controlled sample expansion reduces the accumulation of erroneous pseudo-labels and prevents majority-class dominance during iterative retraining.

The complete architecture of the proposed framework is illustrated in Figure 1. As shown, the model consists of three tightly coupled modules: (i) text preprocessing and feature construction, (ii) structured hybrid feature fusion and classification, and (iii) controlled semi-supervised iterative refinement. The interaction among these modules enables the framework to simultaneously exploit contextual semantics, explicit linguistic cues, and unlabeled supervisory information within a single low-resource learning pipeline.

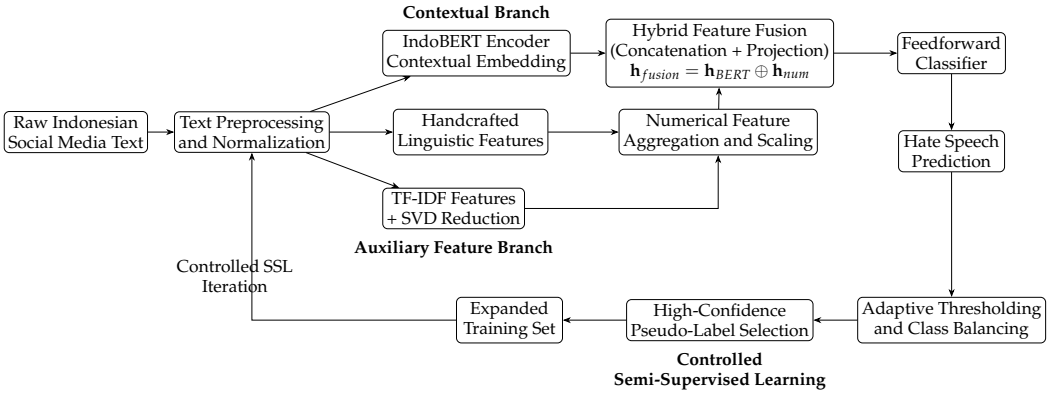


Figure 1. Overview of the proposed hybrid feature-enhanced IndoBERT framework with controlled semi-supervised learning. The model integrates contextual embeddings from IndoBERT with handcrafted linguistic features and TF-IDF–SVD representations through hybrid feature fusion. High-confidence pseudo-labels are selected using adaptive thresholding and class-balanced filtering, enabling iterative expansion of the training data under low-resource conditions.

Formally, the overall prediction function of the proposed architecture is expressed as Equation (1).

$$\hat{y} = f_\theta\big(\phi_{\text{BERT}}(x') \oplus \text{Norm}(\phi_{\text{num}}(x'))\big), \tag{1}$$

where $\phi_{\text{BERT}}(x')$ denotes the contextual embedding extracted from IndoBERT, $\phi_{\text{num}}(x')$ represents the auxiliary numerical feature vector constructed from handcrafted and TF-IDF–SVD descriptors, $\text{Norm}(\cdot)$ denotes feature standardization, $\oplus$ indicates feature concatenation, and f_θ is the downstream nonlinear classifier.

Equation (1) summarizes the central principle of the proposed method: hate speech prediction is determined not by contextual semantic modeling alone but by the joint interaction between contextual, lexical–statistical, and iteratively expanded supervisory information. The detailed construction of each module is presented in the following subsections.

3.2. Text Preprocessing and Auxiliary Feature Construction

Indonesian social media hate speech data are characterized by substantial lexical irregularity, including URL contamination, user mentions, hashtags, orthographic elongation, abbreviated slang, repeated punctuation, and nonstandard abusive expressions [42,43]. Such characteristics introduce severe sparsity into both contextual tokenization and sparse statistical feature extraction [17]. Therefore, before contextual encoding and numerical representation learning are performed, a dedicated preprocessing stage is applied to normalize the raw textual input.

Given an original text sample x, the normalized text is obtained through a composite preprocessing transformation (Equation (2)).

$$x' = \mathcal{P}(x), \tag{2}$$

where $\mathcal{P}(\cdot)$ sequentially applies: (i) URL and mention removal, (ii) hashtag normalization, (iii) punctuation standardization, (iv) repeated-character compression, (v) lowercasing, and (vi) slang lexical normalization using an Indonesian colloquial dictionary.

This normalization step serves two important functions. First, it reduces token fragmentation within the IndoBERT tokenizer by converting orthographically inconsistent forms into more stable lexical units. Second, it decreases sparsity in handcrafted and TF-IDF descriptors by collapsing semantically equivalent noisy variants into unified textual representations. Consequently, both contextual and auxiliary feature branches operate on a significantly cleaner input manifold.

After preprocessing, two groups of auxiliary numerical descriptors are extracted to complement contextual IndoBERT embeddings: handcrafted linguistic indicators and TF-IDF–SVD statistical representations.

3.2.1. Handcrafted Linguistic Indicators

Although transformer encoders are effective in modeling semantic context, they do not explicitly preserve several sparse but highly discriminative hate-speech cues that frequently appear in informal Indonesian discourse. To capture such explicit indicators [87], a handcrafted feature vector is constructed from the normalized text x', as shown in Equation (3).

$$\mathbf{h}_{\text{hand}} = [f_1(x'), f_2(x'), \ldots, f_k(x')], \tag{3}$$

where each $f_i(\cdot)$ denotes a manually designed linguistic statistic.

Handcrafted descriptors include abusive lexicon frequency, punctuation aggression patterns, uppercase emphasis ratio, repeated-character intensity, negation markers, informal intensifier cues, and text-structure indicators. These features provide direct evidence of explicit and stylistic textual aggression that may not be sufficiently emphasized by dense contextual embeddings, particularly when labeled supervision is limited and transformer fine-tuning remains unstable.

To improve reproducibility and interpretability, the handcrafted feature branch is explicitly defined as a compact numerical descriptor consisting of lexicon-based abusive cues, orthographic indicators, punctuation-based patterns, negation markers, informal intensifier cues, and text-structure statistics. Each feature is computed from the normalized text x' after preprocessing, so that informal spelling variants, repeated characters, URLs, mentions, and slang forms are handled consistently before feature extraction. The handcrafted feature vector is intentionally kept low-dimensional to avoid dominating the contextual IndoBERT embedding and the TF-IDF–SVD representation. Table 1 summarizes the handcrafted feature groups, extraction rules, and resulting feature dimensions.

Table 1. Handcrafted linguistic indicators used in the auxiliary feature branch.

Feature Group	Feature	Extraction Rule	Dim.
Lexical hostility	Abusive lexicon count	Number of tokens matching the Indonesian abusive-word lexicon after normalization.	1
Lexical hostility	Abusive lexicon ratio	Abusive lexicon count divided by the total number of tokens in the text.	1
Punctuation aggression	Exclamation density	Number of exclamation marks divided by total character length.	1
Punctuation aggression	Question mark density	Number of question marks divided by total character length.	1
Punctuation aggression	Repeated punctuation pattern	Count of repeated punctuation sequences such as “!!”, “???”, “?!”, or “!?”.	1
Orthographic emphasis	Uppercase ratio	Number of uppercase alphabetic characters divided by total alphabetic characters.	1
Character elongation	Repeated-character intensity	Count of elongated character patterns containing three or more repeated characters.	1
Negation and polarity cue	Negation marker count	Number of Indonesian negation markers such as *tidak*, *tak*, *bukan*, *jangan*, *gak*, or *nggak*.	1

Table 1. *Cont.*

Feature Group	Feature	Extraction Rule	Dim.
Stylistic hostility	Intensifier marker count	Number of informal intensifier or emotionally amplified markers such as *banget*, *parah*, *anjir*, or *woy*.	1
Text structure	Token length	Number of tokens after preprocessing.	1
Text structure	Average token length	Average number of characters per token.	1
Total	–	–	11

Importantly, handcrafted indicators preserve interpretable lexical and stylistic hostility signatures. For example, abusive lexicon counts directly reflect explicit profanity or abusive-word presence, while punctuation repetition, repeated characters, uppercase intensity, and informal intensifiers capture emotionally amplified expressions often associated with offensive or provocative user behavior. All handcrafted indicators are standardized using z-score normalization before being concatenated with the TF-IDF–SVD representation.

3.2.2. TF-IDF–SVD Statistical Representation

In addition to handcrafted linguistic cues, a statistical representation is constructed to preserve global lexical distribution patterns that remain highly informative under sparse supervision [20,21]. Specifically, the normalized text is transformed into a high-dimensional TF-IDF vector [68,88] (Equation (4)).

$$\mathbf{t} = \mathrm{TFIDF}(x'). \tag{4}$$

This sparse vector captures both word-level and character-level frequency salience, allowing the model to retain local lexical patterns, subword hostility fragments, and recurring phrase structures that are common in Indonesian online hate speech.

To ensure reproducibility, the TF-IDF representation was constructed using both word-level and character-level n-grams. Word n-grams were used to capture explicit lexical and phrase-level hostility patterns, while character n-grams were included to preserve subword variations, informal spelling, and orthographic distortions commonly found in Indonesian social media text. The TF-IDF vocabulary was limited to the most informative terms to reduce sparsity and computational overhead, while extremely rare and overly frequent terms were filtered using minimum and maximum document-frequency thresholds. The main TF-IDF–SVD configuration used in the experiments is summarized in Table 2.

However, direct use of high-dimensional TF-IDF vectors creates substantial dimensional imbalance relative to other feature groups [89,90]. Therefore, Singular Value Decomposition (SVD → Equation (5)) is employed to project the sparse representation into a lower-dimensional latent semantic space.

$$\mathbf{h}_{\mathrm{svd}} = \mathbf{U}_r^{\top}\mathbf{t}, \tag{5}$$

where $\mathbf{U}_r$ contains the top-r principal singular directions.

The retained SVD dimension r controls the trade-off between statistical information preservation and compactness. A very small value of r may discard useful lexical-frequency patterns, whereas an excessively large value may increase computational cost and enlarge the auxiliary feature branch. In this study, several candidate dimensions ($r = 50, 100, 200, 300$) were examined in terms of explained variance, and $r = 200$ was selected as a practical trade-off between compactness and statistical information preservation.

Table 2. TF-IDF–SVD configuration used in the auxiliary statistical branch.

Parameter	Setting
Input text	Preprocessed and normalized Indonesian social media text
Corpus size after merging and duplicate removal	14,184 text samples
Word n-gram range	Unigram and bigram $(1, 2)$
Character n-gram range	Character 3-gram to 5-gram $(3, 5)$
Maximum word-level vocabulary size	20,000
Maximum character-level vocabulary size	20,000
Total TF-IDF feature dimension	40,000
Minimum document frequency	2
Maximum document frequency	0.90
TF-IDF normalization	L2 normalization
Dimensionality reduction	Truncated Singular Value Decomposition
Candidate retained SVD dimensions r	50, 100, 200, and 300
Selected retained SVD dimension r	200
Explained variance ratio at $r = 200$	24.72%
Post-SVD scaling	z-score standardization before fusion

The selected TF-IDF representation produced a 40,000-dimensional sparse vector, consisting of 20,000 word-level features and 20,000 character-level features. Character n-grams were included because Indonesian social media hate speech often contains informal spelling, abbreviation, character elongation, and subword-level hostility patterns. Truncated SVD was then applied to obtain a compact dense representation. The retained SVD rank was set to $r = 200$, which preserved 24.72% of the total sparse TF-IDF variance. Although this explained variance ratio is lower than values commonly observed in dense numerical data, it is expected in high-dimensional sparse lexical spaces where discriminative information is distributed across many rare terms and n-gram patterns. Therefore, the role of SVD in this framework is not to preserve all lexical variance but to obtain a compact and less noisy statistical descriptor that can complement IndoBERT embeddings without dominating the fused representation.

As shown in Table 3, increasing r improves the retained statistical variance, from 11.91% at $r = 50$ to 30.19% at $r = 300$. The value $r = 200$ preserves 24.72% of the TF-IDF variance while keeping the auxiliary representation compact. Although $r = 300$ preserves more variance, it also increases the dimensionality of the auxiliary branch and the computational cost. Therefore, $r = 200$ was adopted as the default SVD dimension in the main hybrid model.

Table 3. Explained-variance sensitivity analysis for different retained SVD dimensions.

SVD Dimension r	SVD Output Dimension	Explained Variance Ratio
50	50	11.91%
100	100	17.19%
200	200	24.72%
300	300	30.19%

This transformation preserves the dominant statistical structure of the corpus while suppressing sparse lexical noise and redundant dimensionality. As a result, the TF-IDF–SVD branch provides a compact global descriptor of frequency-based hostility patterns, complementing the local handcrafted indicators and the deep semantic contextual representation.

3.2.3. Auxiliary Numerical Aggregation

The handcrafted vector and TF-IDF–SVD statistical vector are concatenated to form the auxiliary numerical representation (Equation (6)).

$$\mathbf{h}_{\text{num}} = \mathbf{h}_{\text{hand}} \oplus \mathbf{h}_{\text{svd}}. \tag{6}$$

Since handcrafted and statistical features originate from heterogeneous scales and distributions, direct fusion would cause unstable dominance by high-variance components. Therefore, feature standardization is applied (Equation (7)).

$$\mathbf{h}'_{\text{num}} = \text{Norm}(\mathbf{h}_{\text{num}}), \tag{7}$$

where $\text{Norm}(\cdot)$ denotes z-score normalization.

This normalized auxiliary representation serves as the complete explicit lexical–statistical descriptor supplied to the hybrid fusion module. By preserving handcrafted hostility cues together with latent TF-IDF semantic structure, the auxiliary branch ensures that sparse but critical discriminatory information is not lost during transformer-based contextual abstraction.

3.3. Hybrid Feature Fusion and Classification

After contextual and auxiliary numerical representations are constructed, the next critical step is to integrate these heterogeneous feature spaces into a unified discriminative representation. This integration is nontrivial because the two branches encode fundamentally different information modalities. The IndoBERT branch produces dense contextual semantic embeddings that capture long-range linguistic dependencies and sentence-level intent, whereas the auxiliary branch produces sparse-to-compact numerical descriptors that explicitly encode lexical hostility cues and global statistical structure [55,56,61]. Direct use of either branch alone is insufficient: contextual embeddings may suppress sparse explicit hate markers, while numerical descriptors cannot model deep semantic relationships. Therefore, a lightweight hybrid fusion mechanism is introduced to align and jointly exploit both representations.

Let the contextual representation extracted from IndoBERT be denoted by Equation (8),

$$\mathbf{h}_{\text{BERT}} \in \mathbb{R}^{d_b}, \tag{8}$$

and the normalized auxiliary numerical representation be denoted by Equation (9).

$$\mathbf{h}'_{\text{num}} \in \mathbb{R}^{d_n}. \tag{9}$$

The two feature groups are first concatenated to form the raw hybrid vector (Equation (10)).

$$\mathbf{h}_{\text{fusion}} = \mathbf{h}_{\text{BERT}} \oplus \mathbf{h}'_{\text{num}}. \tag{10}$$

Equation (10) represents the first-stage alignment of contextual semantics with explicit lexical–statistical descriptors. Unlike isolated feature usage, this unified representation preserves both high-level semantic intent and sparse discriminative hate-speech indicators within a common feature manifold.

It should be noted that the proposed fusion module is intentionally designed as a lightweight concatenation–projection mechanism rather than as a complex cross-attention or gated interaction architecture. This design choice is motivated by the low-resource nature of the task. When only 5%, 10%, or 20% of the labeled data are available, highly parameterized fusion mechanisms may increase optimization instability, overfitting risk,

and computational cost. Therefore, the concatenation stage is used to preserve all contextual, handcrafted, and statistical information in a single representation, while the subsequent projection layer provides a compact nonlinear transformation that allows cross-feature interaction without introducing excessive additional parameters.

However, simple concatenation alone does not guarantee effective interaction between heterogeneous feature groups. The fused vector still contains independently encoded components that may not yet contribute synergistically to the final decision process. To enable nonlinear cross-feature interaction, the concatenated representation is passed through a trainable projection layer (Equation (11)).

$$\mathbf{z} = \sigma(\mathbf{W}_1 \mathbf{h}_{\text{fusion}} + \mathbf{b}_1), \tag{11}$$

where $\mathbf{W}_1$ and $\mathbf{b}_1$ denote trainable parameters and $\sigma(\cdot)$ is the ReLU activation function.

Compared with more advanced interaction mechanisms such as cross-attention, gated fusion, or adaptive feature weighting, the proposed projection-based fusion is less expressive but more stable and computationally efficient under limited supervision. The purpose of this study is therefore not to claim architectural novelty from the fusion operation itself but to evaluate whether a simple and reproducible fusion module can effectively combine contextual semantic representations, explicit linguistic hostility indicators, and compact statistical descriptors when paired with controlled semi-supervised learning. More sophisticated interaction mechanisms are considered promising future extensions, particularly when larger labeled corpora are available.

This projection serves three simultaneous purposes. First, it performs dimensional adaptation between heterogeneous feature scales. Second, it learns nonlinear dependency interactions between semantic contextual signals and explicit numerical indicators. Third, it suppresses redundant dimensions while emphasizing jointly informative combinations of features that are highly predictive for hate speech discrimination.

In practical terms, the projection layer enables the model to learn patterns such as semantically ambiguous statements that become offensive only when accompanied by explicit identity references, punctuation aggression, or latent lexical hostility signatures captured by the auxiliary branch. Thus, the model does not rely on contextual and numerical features as completely isolated signals but uses the projection layer to support nonlinear alignment between semantic, handcrafted, and statistical information.

The projected latent representation $\mathbf{z}$ is then passed to the final classifier:

$$\hat{y} = \text{Softmax}(\mathbf{W}_2 \mathbf{z} + \mathbf{b}_2), \tag{12}$$

where $\hat{y}$ denotes the predicted class probability distribution over hate speech and non-hate speech labels.

The classification network is optimized using the cross-entropy objective:

$$\mathcal{L}_{ce} = -\sum_{i=1}^{C} y_i \log(\hat{y}_i), \tag{13}$$

where C denotes the number of classes, y_i is the ground-truth class indicator, and $\hat{y}_i$ is the predicted probability.

To further compensate for imbalance between hate speech and non-hate speech categories, class weighting is incorporated into the optimization:

$$\mathcal{L}_{wce} = -\sum_{i=1}^{C} w_i y_i \log(\hat{y}_i), \tag{14}$$

where w_i denotes the inverse-frequency class weight assigned to class i.

This weighted optimization is particularly important because low-resource hate speech corpora often exhibit skewed class distributions, causing conventional optimization to favor the majority class. By amplifying minority-class error contributions, the weighted objective encourages the fused model to maintain balanced decision boundaries rather than maximizing raw accuracy alone.

The complete hybrid decision function of the proposed framework can therefore be summarized as:

$$\hat{y} = f_\theta\left(\phi_{\text{BERT}}(x') \oplus \text{Norm}\left[\phi_{\text{hand}}(x') \oplus \phi_{\text{svd}}(x')\right]\right), \tag{15}$$

where f_θ denotes the trainable nonlinear projection and classification operator.

Figure 2 illustrates this lightweight concatenation–projection fusion process. The upper branch supplies dense contextual semantics, while the lower branch contributes normalized explicit numerical descriptors. Both branches are merged through concatenation, transformed by nonlinear projection, and subsequently mapped into final class probabilities. This architecture ensures that the final prediction is not derived from contextual language understanding alone but from a cooperative interaction between semantic interpretation, lexical hostility evidence, and statistical text structure.

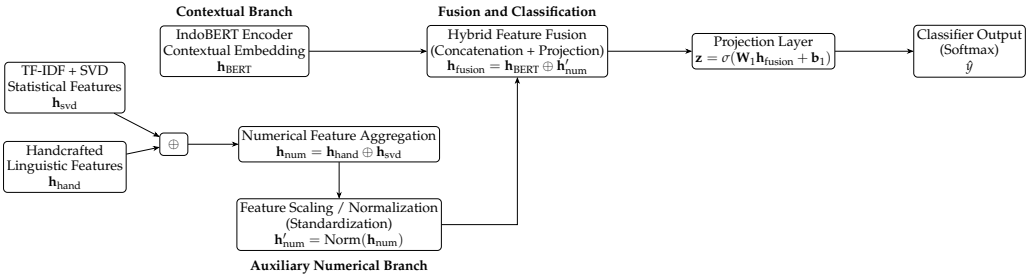


Figure 2. Detailed feature fusion process of the proposed model. Contextual embeddings from IndoBERT are combined with normalized auxiliary numerical features derived from handcrafted linguistic indicators and TF-IDF–SVD representations through a lightweight concatenation–projection fusion mechanism.

3.4. Controlled Semi-Supervised Learning

Although the hybrid feature-enhanced supervised model substantially improves discrimination under limited labeled data, its learning capacity remains constrained by the small size of annotated Indonesian hate speech corpora. A large portion of the merged dataset remains unlabeled during each low-resource split, containing potentially useful lexical and semantic patterns that are not directly exploited by purely supervised optimization. To address this limitation, a controlled semi-supervised learning mechanism is introduced to iteratively assimilate reliable unlabeled samples into the training process.

Unlike conventional pseudo-labeling, which naively assigns labels to all confident predictions and often suffers from confirmation bias [25], the proposed framework employs a reliability-controlled pseudo-label refinement strategy. This strategy is designed to satisfy three simultaneous objectives: (i) suppress noisy pseudo-label injection, (ii) progressively enlarge the effective training distribution, and (iii) maintain class-balanced expansion during iterative learning.

Let the unlabeled sample pool be denoted by Equation (16).

$$\mathcal{D}_U = \{x_u^{(1)}, x_u^{(2)}, \ldots, x_u^{(M)}\}, \tag{16}$$

where M is the number of unlabeled instances available after selecting the small labeled subset.

After training the hybrid supervised classifier in iteration t, the model predicts posterior probabilities for each unlabeled sample (Equation (17)):

$$p^{(t)}(x_u) = P(y = 1|x_u; f_\theta^{(t)}). \tag{17}$$

These posterior probabilities quantify the current confidence of the model regarding hate speech membership. However, directly using all predictions is unsafe because low-confidence samples may introduce substantial semantic noise and reinforce early model bias. Therefore, pseudo-label assignment is performed only under a confidence-gated selection rule (Equation (18)).

$$\tilde{y}^{(t)} = \begin{cases} 1, & p^{(t)}(x_u) \geq \tau_t \\ 0, & p^{(t)}(x_u) \leq 1 - \tau_t \\ \text{discard}, & \text{otherwise} \end{cases} \tag{18}$$

where τ_t denotes the iteration-dependent confidence threshold.

This formulation ensures that only strongly separable unlabeled samples are admitted into the pseudo-labeled set, while ambiguous instances located near the decision boundary are intentionally excluded. As a result, the pseudo-label pool is dominated by samples with high semantic certainty, reducing the probability of propagating mislabeled noise.

To avoid overly rigid sample selection in later iterations, the confidence threshold is not kept constant. Instead, an adaptive threshold relaxation schedule is introduced in Equation (19).

$$\tau_t = \max(\tau_{\min}, \tau_0 - \delta t), \tag{19}$$

where τ_0 is the initial strict confidence threshold, δ is the relaxation coefficient, and $\tau_{\min}$ is the minimum allowable threshold.

The intuition behind Equation (19) is that early pseudo-labeling iterations should remain conservative because the classifier is still imperfect and prone to uncertainty. As iterative retraining improves the model, the threshold can gradually be relaxed to admit a larger set of unlabeled samples without significantly compromising reliability. This produces a curriculum-style expansion of the effective training set: highly certain samples are learned first, while moderately certain samples are incorporated only after the classifier becomes more stable.

A second major risk in pseudo-labeling is class-distribution collapse. In imbalanced hate speech corpora, the model tends to assign more confident predictions to the majority class, causing unlabeled expansion to disproportionately reinforce already dominant labels. To prevent this phenomenon, class-balanced pseudo-label filtering is applied (Equation (20)).

$$|\tilde{\mathcal{D}}_{\text{HS}}^{(t)}| \approx |\tilde{\mathcal{D}}_{\text{NonHS}}^{(t)}|, \tag{20}$$

where $\tilde{\mathcal{D}}_{\text{HS}}^{(t)}$ and $\tilde{\mathcal{D}}_{\text{NonHS}}^{(t)}$ denote the selected pseudo-labeled hate speech and non-hate speech subsets, respectively.

This balancing mechanism ensures that pseudo-label growth does not simply amplify the dominant non-hate class, but instead supplies approximately symmetric supervisory signals to both categories. Consequently, the classifier learns a more stable and less biased decision boundary across iterative refinement.

After confidence filtering and class balancing, the accepted pseudo-labeled set $\tilde{\mathcal{D}}^{(t)}$ is merged into the labeled training pool (Equation (21)).

$$\mathcal{D}_L^{(t+1)} = \mathcal{D}_L^{(t)} \cup \tilde{\mathcal{D}}^{(t)}. \tag{21}$$

The hybrid model is then retrained on the expanded dataset, generating an updated classifier $f_\theta^{(t+1)}$. This process is repeated iteratively until the unlabeled pool is exhausted or the pseudo-label quality gain becomes marginal.

Figure 3 illustrates this reliability-controlled iterative refinement process. Unlabeled data are first scored by the current hybrid classifier and then filtered through adaptive confidence thresholding, followed by class-balanced pseudo-label selection. Only the accepted reliable pseudo-labels are reinjected into the next training cycle. This iterative design differs substantially from naive self-training because pseudo-label admission is explicitly regulated by reliability and distributional constraints rather than by unrestricted confidence alone.

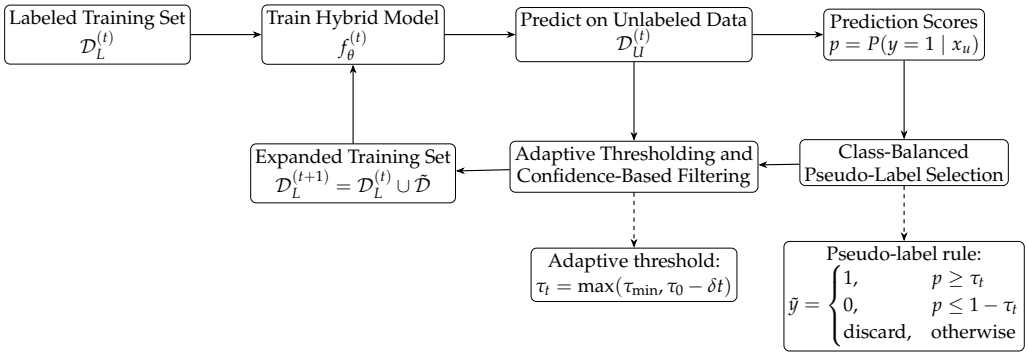


Figure 3. Controlled semi-supervised learning process. Unlabeled samples are predicted by the current hybrid model, filtered using adaptive confidence thresholds, selected with class balancing, and added to the training set for the next training iteration. Solid arrows indicate the main processing flow, whereas dashed arrows connect the filtering stage to the corresponding thresholding and pseudo-labeling rules.

To make the pseudo-labeling process more transparent, the reliability of the selected pseudo-labeled samples is monitored at each SSL iteration. Specifically, four indicators are recorded: pseudo-label acceptance rate, class-wise pseudo-label distribution, estimated pseudo-label noise ratio, and confidence calibration. The acceptance rate measures the proportion of unlabeled samples admitted into the training pool after confidence filtering and class balancing. The class-wise distribution verifies whether pseudo-label expansion remains approximately balanced between hate speech and non-hate speech categories. The pseudo-label noise ratio is estimated on a validation-observable subset by comparing predicted pseudo-labels with available ground-truth annotations when such labels are held out for analysis. Finally, confidence calibration is assessed using Expected Calibration Error (ECE), which evaluates whether high-confidence predictions are consistent with empirical correctness.

$$A^{(t)} = \frac{|\tilde{D}^{(t)}|}{|D_U^{(t)}|}, \tag{22}$$

where $A^{(t)}$ denotes the pseudo-label acceptance rate at SSL iteration t, $|\tilde{D}^{(t)}|$ is the number of accepted pseudo-labeled samples, and $|D_U^{(t)}|$ is the number of remaining unlabeled samples before selection. A very high acceptance rate in early iterations may indicate

overly permissive filtering, whereas a very low acceptance rate may indicate that the threshold is too conservative.

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} |\mathrm{acc}(B_m) - \mathrm{conf}(B_m)|, \tag{23}$$

where B_m denotes the m-th confidence bin, n is the total number of evaluated samples, $\mathrm{acc}(B_m)$ is the empirical accuracy of samples in the bin, and $\mathrm{conf}(B_m)$ is the average predicted confidence. Lower ECE values indicate better calibration, meaning that the model's confidence scores are more reliable for pseudo-label selection. These monitoring indicators are used to support the empirical pseudo-label reliability and calibration analysis reported in Section 5.6.

Overall, the controlled semi-supervised mechanism transforms unlabeled data from a passive unused resource into an actively curated supervisory source. By progressively expanding the labeled manifold with high-confidence and class-balanced pseudo-labels, the framework improves representation robustness, stabilizes decision boundaries, and enhances generalization under severe low-resource supervision.

3.5. Training Strategy and Optimization Protocol

Training the proposed hybrid framework presents several practical challenges beyond standard supervised text classification. First, the low-resource labeled subsets contain only a small fraction of annotated hate speech examples, causing severe class imbalance and unstable gradient updates. Second, the hybrid architecture combines dense transformer embeddings with heterogeneous auxiliary numerical descriptors, which introduces optimization sensitivity during early convergence. Third, the iterative pseudo-label expansion process requires the classifier to maintain sufficiently calibrated confidence estimates so that reliable unlabeled samples can be selected in subsequent semi-supervised iterations. For these reasons, a carefully controlled optimization protocol is employed.

3.5.1. Weighted Objective for Imbalanced Learning

The merged Indonesian hate speech corpus exhibits a nonuniform class distribution between hate speech and non-hate speech instances. Under low-resource sampling, this imbalance becomes even more pronounced because random subset selection may reduce minority hate speech coverage [38,78]. If conventional cross-entropy optimization is applied directly, the classifier tends to minimize loss by favoring the dominant class, resulting in inflated accuracy but unstable Macro-F1.

To counteract this effect, an inverse-frequency weighted cross-entropy is used (Equation (24)).

$$\mathcal{L}_{wce} = -\sum_{i=1}^{C} w_i y_i \log(\hat{y}_i), \tag{24}$$

where w_i is the class weight assigned and is inversely proportional to class frequency.

This weighted objective amplifies the contribution of minority-class misclassification during backpropagation, forcing the model to learn more balanced discriminatory boundaries rather than majority-class memorization. The use of weighted optimization is particularly important because Macro-F1, rather than raw accuracy, serves as the principal evaluation metric of this study.

3.5.2. Balanced Mini-Batch Sampling

In addition to weighted loss, mini-batch construction is controlled using weighted random sampling [91]. During each epoch, samples are drawn with probabilities inversely

proportional to class occurrence, ensuring that hate speech and non-hate speech examples are presented to the optimizer in approximately balanced proportions.

This batch-level balancing reduces the gradient oscillation caused by dominant-class overexposure and stabilizes the representation learning process of both the contextual and auxiliary branches. More importantly, it prevents the pseudo-labeling module from inheriting strongly biased posterior distributions during early supervised training.

3.5.3. Optimization and Convergence Stabilization

The model parameters are optimized using the Adam optimizer due to its adaptive moment estimation capability, which is effective for hybrid architectures involving heterogeneous parameter scales. Let θ denote the trainable parameter set. The parameter update follows Equation (25).

$$\theta_{t+1} = \theta_t - \eta \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon}, \tag{25}$$

where η is the learning rate, and $\hat{m}_t$ and $\hat{v}_t$ are the bias-corrected first and second moment estimates.

Adam is particularly suitable in this framework because the transformer fine-tuning parameters and the auxiliary projection-layer parameters exhibit different gradient magnitudes. Adaptive learning rates, therefore, help maintain synchronized convergence between the contextual encoder and the numerical fusion branch.

To further stabilize training, a linear warm-up and gradual learning-rate decay schedule are employed. Early warm-up prevents abrupt destructive updates on the pre-trained IndoBERT weights, while later decay enables smoother convergence toward locally stable minima.

3.5.4. Validation-Guided Early Stopping

Low-resource supervised training is highly vulnerable to overfitting because the model can quickly memorize sparse lexical patterns in the small labeled subset. This risk is amplified in hybrid fusion architectures due to their increased representational capacity. Therefore, validation-guided early stopping is incorporated.

Let $\mathcal{M}_{val}^{(e)}$ (Equation (26)) denote the validation Macro-F1 score in epoch e. Training is terminated when no improvement is observed for a predefined patience interval.

$$\mathcal{M}_{val}^{(e)} \leq \max_{j<e} \mathcal{M}_{val}^{(j)} \quad \text{for } P \text{ consecutive epochs.} \tag{26}$$

The checkpoint achieving the highest validation Macro-F1 is preserved as the supervised base model for subsequent pseudo-label refinement. This criterion is preferred over validation accuracy because Macro-F1 better reflects minority hate-speech sensitivity.

3.5.5. Iterative Semi-Supervised Retraining Protocol

After the initial supervised model converges, the controlled semi-supervised module generates a pseudo-labeled expansion set according to Section 3.4. The model is then retrained on the updated labeled pool rather than merely fine-tuned for a few gradient steps. This full retraining strategy is important because it allows newly admitted pseudo-labeled samples to participate uniformly in the optimization process and reduces overdependence on the initial supervised decision boundary.

At each semi-supervised iteration, the same weighted objective, balanced mini-batch sampling, and validation-guided stopping criteria are preserved. This ensures optimization consistency across all pseudo-label refinement cycles and prevents divergence between supervised and semi-supervised learning dynamics.

Overall, this training protocol provides four layers of optimization control: imbalance-aware weighted learning, batch-level class stabilization, adaptive gradient convergence, and validation-regulated iterative retraining. These mechanisms collectively ensure that the proposed hybrid framework remains numerically stable, resistant to overfitting, and sufficiently calibrated for reliable pseudo-label assimilation under severe low-resource supervision.

3.6. Unified Processing Pipeline and Algorithmic Workflow

The complete operational workflow of the proposed framework can now be summarized as a sequential hybrid learning pipeline that integrates supervised feature fusion and reliability-controlled semi-supervised refinement into a single end-to-end process.

Given the original merged Indonesian hate speech corpus, the dataset is first divided into a small labeled subset $\mathcal{D}_L^{(0)}$ and a large unlabeled subset $\mathcal{D}_U$ according to the designated low-resource supervision ratios (5%, 10%, and 20%). All textual samples are subsequently passed through the preprocessing transformation $\mathcal{P}(\cdot)$ to remove lexical noise, normalize slang forms, and reduce orthographic variability.

For each normalized text sample, the contextual branch extracts IndoBERT embeddings $\phi_{\text{BERT}}(x')$, while the auxiliary branch computes handcrafted linguistic indicators $\phi_{\text{hand}}(x')$ and TF-IDF–SVD statistical descriptors $\phi_{\text{svd}}(x')$. These auxiliary descriptors are aggregated, normalized, and concatenated with contextual embeddings to produce the hybrid representation (Equation (27)).

$$\mathbf{h}_{\text{fusion}} = \phi_{\text{BERT}}(x') \oplus \text{Norm}\left[\phi_{\text{hand}}(x') \oplus \phi_{\text{svd}}(x')\right]. \tag{27}$$

The fused vector is then passed through the nonlinear projection-classification module and optimized under weighted imbalance-aware learning to obtain the initial supervised classifier $f_\theta^{(0)}$.

Once the supervised model converges, the unlabeled subset is evaluated by the classifier to obtain posterior confidence scores. High-confidence unlabeled samples satisfying the adaptive confidence gate and class-balance criteria are selected as pseudo-labeled candidates and merged into the labeled set. The hybrid classifier is then retrained on the expanded labeled pool, yielding an updated model $f_\theta^{(1)}$.

This pseudo-label refinement process is repeated iteratively, allowing the effective labeled manifold to grow from a very small supervision seed toward a broader hybrid-supervised representation space. Throughout each cycle, the model progressively acquires additional lexical hostility patterns, semantic sentence structures, and minority-class decision evidence that were not fully captured during the initial supervised stage.

The complete computational workflow therefore consists of six tightly coupled stages:

1. Low-resource labeled/unlabeled split generation;
2. Preprocessing and lexical normalization;
3. Contextual and auxiliary feature extraction;
4. Hybrid feature fusion and supervised optimization;
5. Confidence-controlled pseudo-label selection;
6. Iterative semi-supervised retraining.

This sequential formulation ensures that the proposed framework does not merely perform static supervised classification, but instead executes a progressive representation expansion process in which contextual semantics, explicit linguistic evidence, and curated unlabeled supervisory signals are jointly integrated.

Algorithm 1 summarizes the complete operational procedure of the proposed method.

Algorithm 1 Unified Hybrid IndoBERT with Controlled Semi-Supervised Learning.

Require: Full corpus $\mathcal{D}$, labeled ratio ρ, initial threshold τ_0, minimum threshold $\tau_{\min}$, decay rate δ, maximum SSL iterations T

Ensure: Final trained classifier $f_\theta^{(T)}$

Split $\mathcal{D}$ into labeled subset $\mathcal{D}_L^{(0)}$ and unlabeled subset $\mathcal{D}_U$ using ratio ρ
for each text sample x in $\mathcal{D}_L^{(0)} \cup \mathcal{D}_U$ **do**
    Apply preprocessing: $x' = \mathcal{P}(x)$
    Extract contextual embedding $\phi_{\text{BERT}}(x')$
    Extract handcrafted features $\phi_{\text{hand}}(x')$
    Extract TF-IDF–SVD features $\phi_{\text{svd}}(x')$
end for
Construct fused hybrid representation:

$$\mathbf{h}_{\text{fusion}} = \phi_{\text{BERT}}(x') \oplus \text{Norm}\left[\phi_{\text{hand}}(x') \oplus \phi_{\text{svd}}(x')\right]$$

Train initial hybrid classifier $f_\theta^{(0)}$ on $\mathcal{D}_L^{(0)}$ using weighted cross-entropy and balanced mini-batch sampling
for $t = 0$ to $T - 1$ **do**
    Predict posterior probabilities for all $x_u \in \mathcal{D}_U$:

$$p^{(t)}(x_u) = P(y = 1 | x_u; f_\theta^{(t)})$$

    Update adaptive threshold:

$$\tau_t = \max(\tau_{\min}, \tau_0 - \delta t)$$

    Select high-confidence pseudo-labels using Equation (18)
    Apply class-balanced filtering to selected pseudo-labeled set $\tilde{\mathcal{D}}^{(t)}$
    Update labeled pool:

$$\mathcal{D}_L^{(t+1)} = \mathcal{D}_L^{(t)} \cup \tilde{\mathcal{D}}^{(t)}$$

    Remove selected samples from unlabeled pool $\mathcal{D}_U$
    Retrain hybrid classifier $f_\theta^{(t+1)}$ on updated $\mathcal{D}_L^{(t+1)}$
end for
return Final classifier $f_\theta^{(T)}$

4. Experiments

This section presents the complete experimental protocol used to evaluate the proposed hybrid feature-enhanced IndoBERT framework under low-resource Indonesian hate speech detection settings. Particular emphasis is placed on corpus construction, low-resource supervision simulation, comparative baseline implementation, reproducibility settings, and metric selection to ensure that the subsequent performance analysis is conducted under a rigorous and fair evaluation environment.

4.1. Corpus Construction and Low-Resource Simulation Protocol

To comprehensively evaluate hate speech detection behavior across diverse Indonesian online linguistic patterns, a merged corpus was constructed from three publicly available benchmark datasets: IDHSD, HS-572, and RE. These corpora contain Indonesian user-generated social media texts manually annotated for binary hate speech classification. The use of multiple data sources is important because each corpus reflects different lexical characteristics, annotation distributions, and informal communication styles, thus reducing the risk of domain-specific overfitting.

Before merging, duplicate samples, inconsistent labels, and noisy malformed entries were removed. Label harmonization was then performed to map all annota-

tions into a unified binary taxonomy consisting of hate speech and non-hate speech classes. This produces a heterogeneous but coherent benchmark corpus suitable for robust low-resource evaluation.

Figure 4 presents the final class-label distribution of the merged dataset. As observed, the corpus exhibits a moderate imbalance toward the non-hate speech category. This imbalance reflects realistic social media moderation conditions and motivates the use of weighted optimization and balanced pseudo-label selection in the proposed framework.

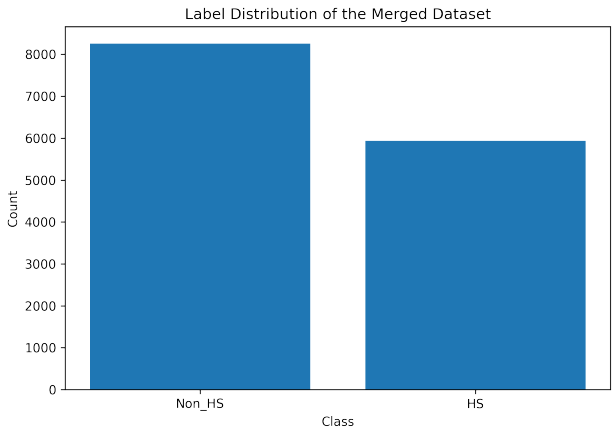


Figure 4. Distribution of hate speech and non-hate speech labels in the merged Indonesian corpus. The corpus exhibits a moderate class imbalance, motivating imbalance-aware optimization.

Figure 5 illustrates the contribution of each original source corpus. The RE dataset contributes the largest volume of samples, while IDHSD and HS-572 provide additional lexical and annotation diversity. This heterogeneous composition ensures that the evaluation is not confined to a single annotation style or domain-specific language pattern.

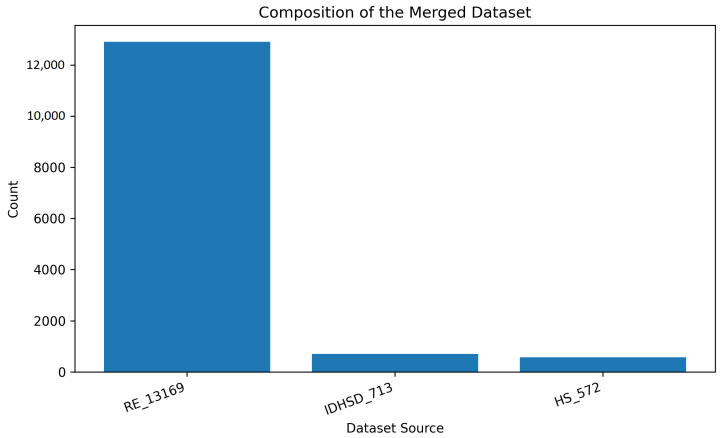


Figure 5. Contribution of each public Indonesian hate speech dataset to the merged benchmark corpus.

To simulate practical annotation scarcity, three low-resource supervision regimes are constructed using labeled fractions of 5%, 10%, and 20%. For each regime, stratified random sampling is applied to preserve the hate speech/non-hate speech ratio in the labeled subset. The remaining samples are treated as unlabeled data and reserved exclusively for semi-supervised pseudo-label refinement.

This design yields an intentionally severe supervision bottleneck in which only a small annotated seed is available to initialize the classifier, closely resembling realistic low-resource moderation settings where manual labeling is expensive and slow.

Figure 6 visualizes the labeled and unlabeled sample proportions under the three supervision scenarios. As the labeled fraction increases, the supervised seed becomes more informative while the unlabeled reservoir decreases correspondingly, allowing systematic analysis of how the proposed framework behaves across different levels of annotation scarcity.

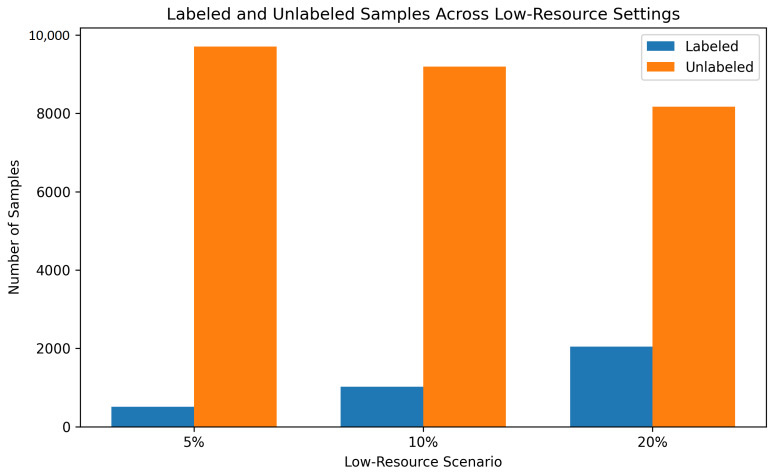


Figure 6. Low-resource supervision simulation under 5%, 10%, and 20% labeled settings. The remaining corpus is reserved as unlabeled data for controlled semi-supervised learning.

To ensure fairness, all baseline models, ablation variants, and representative comparative architectures are trained under identical low-resource splits. Thus, every result reported in the subsequent section reflects the same supervision constraint rather than incomparable data availability.

4.2. Baseline and Comparative Models

To rigorously assess the effectiveness of the proposed framework, a wide range of representative baseline models were implemented under the same merged corpus, identical preprocessing pipeline, identical low-resource split protocol, and identical evaluation metrics. The comparative baselines were selected to cover four major methodological families: (i) classical sparse lexical machine learning models, (ii) shallow neural sequence models, (iii) pretrained transformer-based contextual models, and (iv) conventional semi-supervised self-training approaches.

4.2.1. Classical Sparse Lexical Baselines

Classical machine learning baselines were included to evaluate how far traditional feature-engineered classifiers remain competitive under severe annotation scarcity. These models are particularly important because sparse lexical representations often provide surprisingly stable performance when deep contextual models are under-trained.

The following classical baselines were implemented:

- SVM + TF-IDF: Linear Support Vector Machine trained on TF-IDF unigram and bigram features;
- Logistic Regression + TF-IDF: Logistic classifier using the same sparse lexical representation;
- Random Forest + n-gram: Ensemble tree classifier trained on sparse n-gram count vectors;
- RFDT + Linguistic Features: A random-forest decision-tree configuration trained exclusively on handcrafted linguistic indicators.

These baselines provide reference points for evaluating the contribution of explicit sparse lexical and interpretable feature engineering without contextual transformer representations.

4.2.2. Shallow Neural Text Models

To assess whether lightweight deep architectures can effectively exploit sequential context under low-resource conditions, two standard neural text classifiers were also implemented:

- CNN: A one-dimensional convolutional text classifier with max-pooling over token embeddings;
- LSTM: A recurrent long short-term memory encoder followed by dense classification.

These models capture local and sequential contexts more effectively than classical sparse lexical baselines but do not benefit from large-scale pretrained language representations.

4.2.3. Pretrained Transformer Baselines

Because the proposed method is built upon IndoBERT, several transformer baselines were implemented to isolate the true contribution of feature fusion and controlled semi-supervised learning beyond contextual fine-tuning alone.

The transformer comparative models include:

- IndoBERT: Standard supervised fine-tuning of `indobenchmark/indobert-base-p1`;
- mBERT: Multilingual BERT fine-tuned on the same Indonesian corpus;
- XLM-RoBERTa: Multilingual robustly optimized transformer baseline;
- IndoBERTweet: Indonesian social-media-adapted pretrained transformer.

These baselines allow evaluation of whether the proposed gains originate merely from pretrained contextual semantics or from the additional hybrid feature engineering introduced in this work.

4.2.4. Semi-Supervised Comparative Baseline

To specifically validate the effectiveness of the proposed controlled pseudo-labeling mechanism, a conventional self-training baseline was implemented:

- Naive Pseudo-labeling: Unlabeled samples with fixed high-confidence predictions are directly added to the training set without adaptive threshold relaxation or class-balanced filtering.

This baseline serves as a direct comparison for quantifying the benefit of reliability-controlled pseudo-label admission versus unrestricted confidence-based self-training.

4.2.5. Internal Ablation Variants

In addition to external baselines, several internal ablation variants were constructed to isolate the contribution of each component within the proposed framework:

- IndoBERT only;
- IndoBERT + Handcrafted Features;
- IndoBERT + TF-IDF–SVD;
- Full Hybrid Fusion;
- Full Hybrid Fusion + Controlled SSL (proposed final model).

These ablation variants enable systematic analysis of contextual semantics, auxiliary numerical descriptors, and semi-supervised refinement independently.

Overall, the comparative benchmark therefore spans conventional lexical learning, shallow neural sequence modeling, transformer contextual fine-tuning, naive self-training,

and internal hybrid ablation analysis. Such a broad evaluation space ensures that the reported gains of the proposed framework are not measured against a narrow or artificially weak set of baselines but against a representative spectrum of low-resource hate speech detection strategies.

4.3. Implementation Details and Hyperparameter Settings

To improve experimental reproducibility, the main implementation details and hyperparameter settings used in this study are summarized in Table 4. All experiments were implemented in Python version 3.9.5 using PyTorch version 2.8.0 with CUDA 12.8 support and the Hugging Face Transformers library version 4.57.6. The IndoBERT encoder was initialized from the `indobenchmark/indobert-base-p1` checkpoint. Input texts were tokenized using the corresponding IndoBERT tokenizer with a maximum sequence length of 128 tokens. The model was trained using the AdamW optimizer with a learning rate of 2×10^{-5}, a mini-batch size of 16, and a linear learning-rate scheduler with warm-up. Gradient clipping was applied with a maximum norm of 1.0 to stabilize fine-tuning.

Table 4. Implementation details and hyperparameter settings used in the proposed framework.

Component	Setting
Programming framework	Python, PyTorch, Hugging Face Transformers
Pretrained language model	`indobenchmark/indobert-base-p1`
Tokenizer	IndoBERT tokenizer
Maximum sequence length	128 tokens
Batch size	16
Training epochs	3 epochs
Optimizer	AdamW
Learning rate	2×10^{-5}
Learning-rate scheduler	Linear schedule with warm-up
Warm-up proportion	10% of total training steps
Loss function	Weighted cross-entropy loss
Gradient clipping	Maximum norm = 1.0
Dropout rate	0.20
Auxiliary projection layer	Linear layer with 128 hidden units, ReLU activation, and dropout
Handcrafted feature dimension	9
TF-IDF word n-gram range	(1, 2)
TF-IDF character n-gram range	(3, 5)
Maximum TF-IDF vocabulary	20,000 word-level + 20,000 character-level features
Observed TF-IDF feature dimension	25,033 features under the 20% labeled-data setting
SVD dimension	200
Auxiliary feature dimension	211
Semi-supervised thresholds	$\tau_0 = 0.95$, $\tau_{\min} = 0.85$, $\Delta = 0.05$
SSL iterations	3 iterations
Pseudo-label selection	Confidence-based filtering with class-balanced HS:Non-HS quota
Random seed	42
Hardware	HPC node equipped with 8× NVIDIA A100-SXM4-40GB GPUs; each reported run used CUDA acceleration on a single GPU

For the hybrid model, the auxiliary feature vector consisted of 11 handcrafted linguistic indicators and 200 TF-IDF–SVD components, resulting in a 211-dimensional auxiliary representation. The handcrafted indicators include lexicon-based abusive-word cues as well as punctuation, orthographic, negation, intensifier, and text-structure features. This auxiliary vector was projected into a 128-dimensional dense representation before being concatenated with the IndoBERT contextual embedding. For the controlled semi-supervised learning setting, pseudo-label selection started from a strict confidence threshold of $\tau_0 = 0.95$ and

was gradually relaxed to $\tau_{\min} = 0.85$ with a decrement of $\Delta = 0.05$ across three iterations. At each iteration, pseudo-labeled samples were selected using class-balanced quotas to reduce majority-class dominance during unlabeled-data expansion. All reported training-time, memory, and inference-latency measurements correspond to the same experimental environment and are intended to support relative efficiency comparison among the evaluated variants.

4.4. Evaluation Metrics

To comprehensively evaluate the performance of the proposed framework under low-resource and moderately imbalanced hate speech classification, multiple complementary evaluation metrics are employed. These metrics are selected to assess not only overall correctness but also balanced class discrimination capability and probabilistic ranking quality [22].

4.4.1. Accuracy

Overall classification accuracy is computed as Equation (28).

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}, \tag{28}$$

where TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively.

Accuracy provides a general measure of prediction correctness over the entire test set. However, because hate speech detection typically involves unequal class distributions, this metric alone is insufficient to characterize balanced model behavior.

4.4.2. Precision and Recall

Precision evaluates the reliability of positive hate speech predictions (Equation (29)),

$$\text{Precision} = \frac{TP}{TP + FP}, \tag{29}$$

while recall measures the sensitivity of the model in retrieving actual hate speech instances (Equation (30)).

$$\text{Recall} = \frac{TP}{TP + FN}. \tag{30}$$

These two metrics are particularly relevant because hate speech moderation requires controlling both types of error. Excessive false positives can incorrectly suppress normal content, whereas excessive false negatives can fail to detect harmful abusive expressions.

4.4.3. Macro-F1 Score

To balance precision and recall, the F1-score is computed as Equation (31).

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{31}$$

Because the merged dataset exhibits moderate imbalance, the Macro-F1 score is used as the main evaluation metric (Equation (32)).

$$\text{Macro-F1} = \frac{1}{C} \sum_{c=1}^{C} \text{F1}_c, \tag{32}$$

where C denotes the number of classes.

Macro-F1 assigns equal importance to both hate speech and non-hate speech classes regardless of sample frequency. This makes it more informative than raw accuracy for assessing balanced discrimination under low-resource supervision.

4.4.4. Receiver Operating Characteristic and ROC-AUC

In addition to hard-label classification metrics, probabilistic ranking behavior is evaluated using the Receiver Operating Characteristic (ROC) curve. The true positive rate and false positive rate are defined as Equation (33).

$$\text{TPR} = \frac{TP}{TP + FN}, \qquad \text{FPR} = \frac{FP}{FP + TN}. \tag{33}$$

The Area Under the ROC Curve (ROC-AUC) is computed as Equation (34).

$$\text{ROC-AUC} = \int_0^1 \text{TPR}(\text{FPR})\, d(\text{FPR}). \tag{34}$$

ROC-AUC is particularly important in this study because the proposed semi-supervised mechanism utilizes confidence-based pseudo-label selection. Strong probabilistic separation indicates that the model can assign more reliable confidence scores when selecting pseudo-labeled unlabeled samples.

4.4.5. Metric Selection Rationale

Among all reported metrics, Macro-F1 is used as the primary comparative criterion because it best reflects balanced class sensitivity under imbalanced low-resource classification. Accuracy is reported as a secondary global correctness measure, precision and recall are used to analyze positive prediction reliability and sensitivity, and ROC-AUC is employed to evaluate confidence ranking quality.

The use of these complementary metrics ensures that the proposed framework is evaluated from multiple perspectives, including balanced class discrimination, probabilistic separability, and practical hate speech detection robustness.

5. Results and Discussion

This section presents a comprehensive evaluation of the proposed hybrid feature-enhanced IndoBERT framework under low-resource Indonesian hate speech detection scenarios. Experiments are conducted using three labeled data settings (5%, 10%, and 20%) to simulate varying levels of supervision, reflecting realistic conditions where annotated data are limited. The evaluation focuses on comparing the proposed hybrid model with baseline and ablation variants, including IndoBERT-only, IndoBERT combined with handcrafted features, IndoBERT combined with TF-IDF–SVD features, and the full hybrid model with and without controlled semi-supervised learning.

Performance is assessed using accuracy, Macro-F1, Macro-Precision, Macro-Recall, and ROC-AUC metrics, with Macro-F1 serving as the primary evaluation measure due to class imbalance in hate speech datasets. The experiments aim to analyze not only the overall performance of the proposed method but also the contribution of individual feature groups and the effectiveness of the controlled semi-supervised learning strategy.

The results are organized into five analyses. First, internal model variants are examined to evaluate the contribution of contextual, handcrafted, statistical, and semi-supervised components. Second, the proposed model is compared with representative state-of-the-art baselines under the same experimental protocol. Third, classification behavior is analyzed using a confusion matrix and ROC diagnostics. Fourth, the controlled semi-supervised

learning mechanism is examined in relation to naive pseudo-labeling. Finally, the discussion synthesizes the main findings, limitations, and practical implications.

5.1. Performance Trend Analysis Across Internal Model Variants

This subsection investigates the performance behavior of the proposed framework and its internal ablation variants under progressively increasing low-resource supervision levels. Unlike the broader benchmark comparison presented in the subsequent subsection, the objective here is to analyze how individual feature groups and the controlled semi-supervised learning mechanism influence the learning dynamics of the proposed architecture. Five internal configurations are considered: (i) IndoBERT only, (ii) IndoBERT combined with handcrafted linguistic features, (iii) IndoBERT combined with TF-IDF–SVD statistical features, (iv) the full hybrid feature framework, and (v) the full hybrid framework integrated with controlled semi-supervised learning (SSL). The evaluation is conducted under 5%, 10%, and 20% labeled data settings using Accuracy, Macro-F1, Precision, Recall, and ROC-AUC, with Macro-F1 treated as the primary indicator due to class imbalance.

Figure 7 presents the Macro-F1 trajectory across the three supervision levels. Several notable patterns can be observed. First, all IndoBERT-based configurations maintain relatively strong performance even under severe label scarcity, confirming the robustness of contextual transformer embeddings for low-resource Indonesian hate speech detection. However, the contribution of auxiliary feature groups is not uniform across supervision levels. Under the 5% labeled setting, the IndoBERT + TF-IDF–SVD configuration achieves the highest Macro-F1 score among all internal variants, reaching 0.7990. This result indicates that compact statistical representations derived from global lexical distributions provide stable complementary information when contextual transformer representations are still under-constrained by limited labeled supervision.

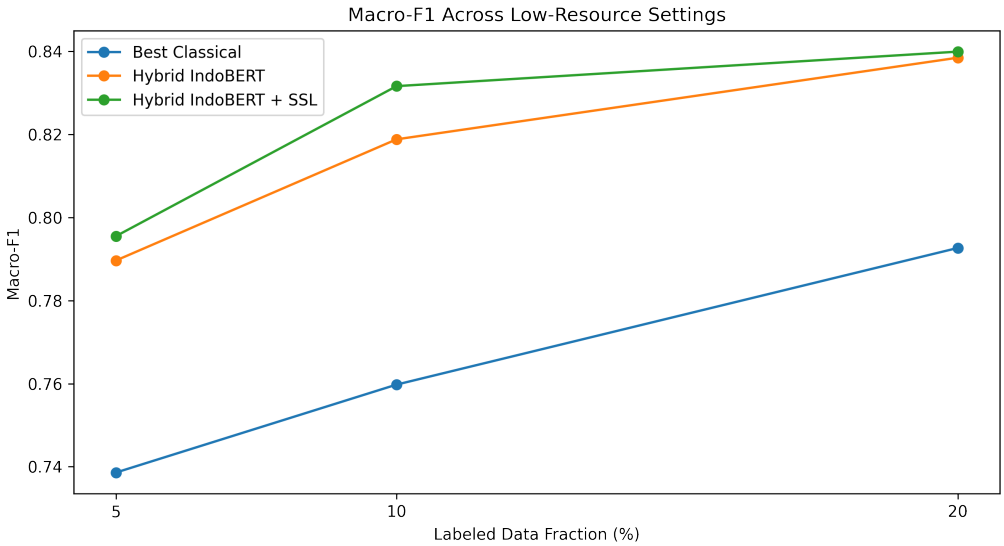


Figure 7. Macro-F1 performance across different label fractions (5%, 10%, and 20%). The statistical TF-IDF–SVD augmentation provides the strongest benefit under extreme low-resource conditions, whereas the full Hybrid + SSL framework becomes dominant as the supervision level increases.

As the labeled fraction increases, the relative contribution of richer feature fusion becomes more apparent. At the 10% supervision level, the Full Hybrid + SSL configuration becomes the best-performing variant, with a Macro-F1 score of 0.8232, and this superiority becomes even more pronounced at the 20% setting, where the score rises to 0.8633. This progressive improvement suggests that more complex interactions among contextual embeddings, handcrafted linguistic cues, and statistical representations require a suffi-

ciently representative labeled subset before their complementarity can be fully exploited. In other words, hybrid feature integration is not uniformly advantageous at all data scales but becomes increasingly effective once the model is able to learn more reliable semantic alignment from the labeled data.

A similar trend is reflected in the accuracy analysis shown in Figure 8. Under the 5% labeled regime, the IndoBERT + TF-IDF–SVD variant remains the most stable configuration, while the Full Hybrid + SSL framework gradually overtakes all other variants with 10% and 20% labeled data. This confirms that statistical features primarily contribute to early-stage robustness, whereas the semi-supervised hybrid architecture contributes more strongly to downstream generalization once a minimally stable supervised backbone has been established.

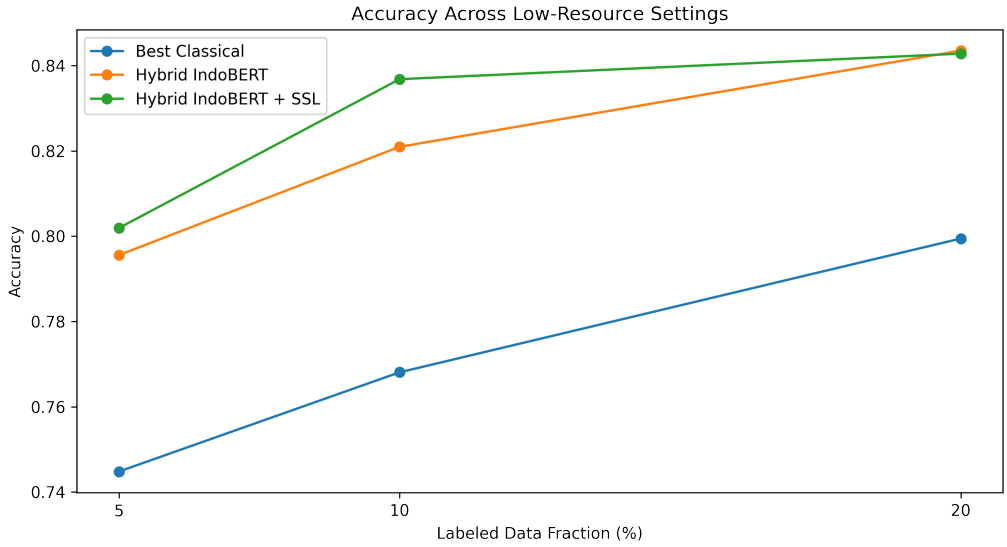


Figure 8. Accuracy comparison across internal model variants under varying labeled data fractions. The Full Hybrid + SSL framework shows the strongest convergence as labeled supervision becomes more representative.

The ROC-AUC behavior in Figure 9 provides additional evidence regarding probabilistic class separability. Although all internal variants show improved ranking behavior as the amount of labeled data increases, the Full Hybrid + SSL model consistently achieves the strongest AUC progression, culminating in 0.9334 at the 20% labeled setting. This indicates that the controlled pseudo-labeling mechanism does not merely improve threshold-specific classification accuracy, but also refines the global ordering of positive and negative confidence distributions, resulting in stronger class separability.

Figure 10 further summarizes the grouped Macro-F1 behavior across all internal variants. The figure clearly shows that handcrafted linguistic features contribute only marginally relative to the IndoBERT baseline, whereas TF-IDF–SVD features provide a stronger complementary signal under the lowest supervision level. However, once pseudo-label refinement is introduced into the complete hybrid framework, the Full Hybrid + SSL variant becomes the most balanced and scalable performer. This indicates that the principal gains do not originate from any single handcrafted feature source but rather from the controlled interaction between heterogeneous feature groups and carefully regulated unlabeled data expansion.

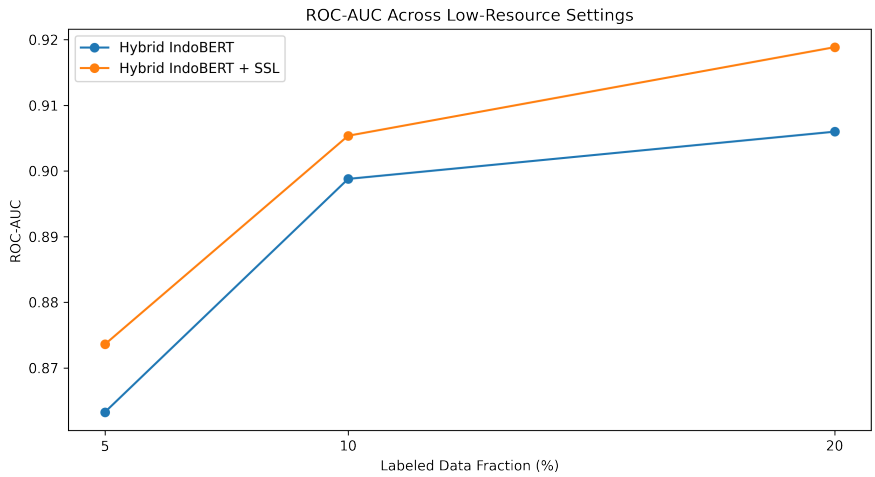


Figure 9. ROC-AUC comparison across internal model variants. The Full Hybrid + SSL framework achieves the strongest probabilistic discrimination under moderate supervision.

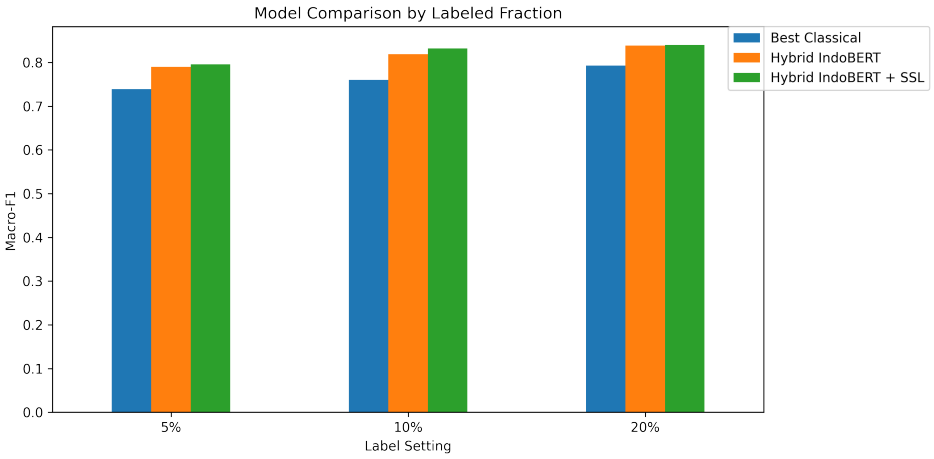


Figure 10. Grouped Macro-F1 comparison across internal model variants. The figure highlights the supervision-dependent transition from TF-IDF–SVD dominance at 5% to Hybrid + SSL dominance at 10% and 20%.

The relative improvement behavior is further illustrated in Figure 11, where performance gains are measured against the weakest internal configuration. The greatest relative gains are observed under the 5% labeled setting, confirming that feature augmentation has the highest marginal utility when supervision is severely constrained. As the labeled fraction increases, the absolute performance still improves, but the relative gap between variants becomes more compressed, indicating that the model gradually approaches a more stable supervised regime where improvements are increasingly driven by refinement rather than fundamental representational rescue.

For completeness, Table 5 summarizes the quantitative performance of all internal model variants. Several important observations emerge. First, the IndoBERT baseline alone already establishes a strong contextual foundation across all label fractions. Second, handcrafted linguistic features introduce only minor fluctuations, suggesting that explicit lexical indicators alone are insufficient to substantially reshape transformer-level representations. Third, TF-IDF–SVD features provide the strongest isolated complementary signal under the 5% regime. Finally, the integration of all feature groups together with controlled semi-supervised learning yields the strongest and most consistent performance once the labeled subset becomes sufficiently representative.

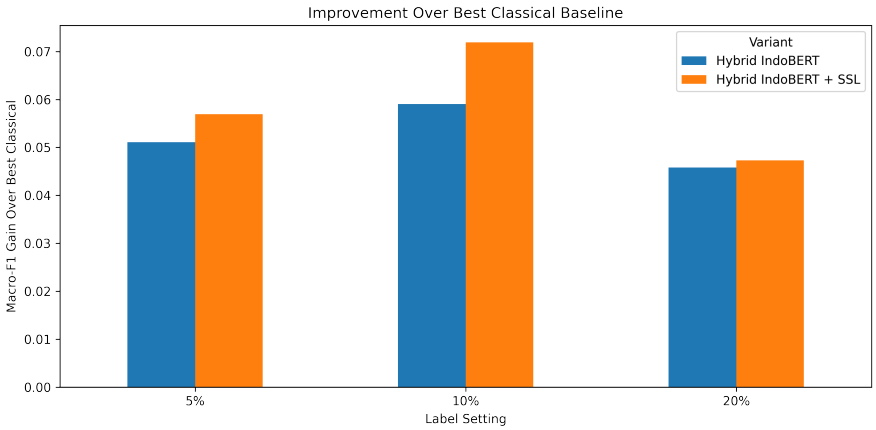


Figure 11. Relative Macro-F1 gain across internal model variants. The largest marginal gains occur under the most constrained supervision setting, highlighting the importance of feature augmentation in severe low-resource learning.

Table 5. Quantitative summary of internal ablation variants across low-resource supervision levels.

Label	Variant	Accuracy	Macro-F1	Precision	Recall	ROC-AUC
5%	IndoBERT only	0.7970	0.7943	0.7930	0.7998	0.8759
	IndoBERT + Handcrafted	0.7963	0.7933	0.7918	0.7981	0.8567
	IndoBERT + TF-IDF–SVD	0.8044	0.7990	0.7990	0.7990	0.8772
	Full Hybrid	0.7956	0.7897	0.7901	0.7893	0.8632
	Full Hybrid + SSL	0.8040	0.7946	0.8037	0.7901	0.8685
10%	IndoBERT only	0.8199	0.8154	0.8147	0.8162	0.8923
	IndoBERT + Handcrafted	0.8238	0.8184	0.8195	0.8174	0.8904
	IndoBERT + TF-IDF–SVD	0.8269	0.8222	0.8222	0.8223	0.8881
	Full Hybrid	0.8231	0.8210	0.8197	0.8275	0.8990
	Full Hybrid + SSL	0.8287	0.8232	0.8249	0.8218	0.8987
20%	IndoBERT only	0.8393	0.8333	0.8372	0.8306	0.9102
	IndoBERT + Handcrafted	0.8364	0.8319	0.8321	0.8316	0.9065
	IndoBERT + TF-IDF–SVD	0.8340	0.8307	0.8289	0.8335	0.9111
	Full Hybrid	0.8340	0.8304	0.8289	0.8326	0.9076
	Full Hybrid + SSL	0.8654	0.8633	0.8423	0.8450	0.9334

Overall, the evidence from the internal variant analysis demonstrates that the effectiveness of each feature component is strongly supervision-dependent. Under extremely constrained labeled regimes, compact statistical augmentation provides the most stable auxiliary signal. As the supervision level becomes moderately richer, the full interaction among contextual embeddings, heterogeneous auxiliary features, and controlled semi-supervised learning becomes the dominant source of improvement. This supervision-dependent transition forms the central empirical motivation for the broader representative baseline comparison presented in the next subsection.

5.2. Comparison with Representative State-of-the-Art Baselines

Although the previous subsection analyzed the internal contribution of each architectural component, a broader evaluation is necessary to determine whether the proposed framework remains competitive compared to representative low-resource hate speech detection baselines spanning multiple methodological families. To this end, a comprehensive benchmark was conducted under the same merged Indonesian hate speech corpus, identical preprocessing pipeline, identical train/validation/test split strategy, and the same low-resource supervision protocol. This controlled evaluation is critical because compar-

isons based only on literature are often affected by dataset heterogeneity, inconsistent preprocessing, and incompatible evaluation metrics. By contrast, all baselines in this study are trained and evaluated under exactly the same conditions, allowing a more rigorous and directly interpretable assessment.

The benchmark includes four categories of representative competitors: (i) classical sparse lexical models, including SVM + TF-IDF, Logistic Regression + TF-IDF, Random Forest + n-gram, and RFDT with handcrafted linguistic features; (ii) shallow neural architectures, namely CNN and LSTM; (iii) pretrained transformer baselines, including IndoBERT, mBERT, XLM-RoBERTa, and IndoBERTweet; and (iv) a conventional naive pseudo-labeling strategy using IndoBERT without adaptive threshold regulation or class-balanced pseudo-label control. Together, these baselines cover the principal methodological families commonly used in low-resource text classification and hate speech detection.

Table 6 presents the complete Macro-F1 and ROC-AUC comparison across all methods under 5%, 10%, and 20% labeled data settings.

Table 6. Comparison with representative baselines and state-of-the-art model families under the same experimental protocol.

Method	5% F1	10% F1	20% F1	5% AUC	10% AUC	20% AUC
SVM + TF-IDF	0.7589	0.7849	0.7964	0.8378	0.8612	0.8815
Logistic Regression + TF-IDF	0.7597	0.7763	0.7894	0.8346	0.8521	0.8767
Random Forest + n-gram	0.7384	0.7640	0.7894	0.8119	0.8413	0.8655
RFDT + Linguistic Features	0.6335	0.6657	0.7002	0.6325	0.6654	0.7038
CNN	0.6731	0.7402	0.7600	0.7377	0.8207	0.8404
LSTM	0.6584	0.6851	0.7244	0.7046	0.7496	0.7989
IndoBERT	0.7799	0.8030	0.8285	0.8654	0.8837	0.9095
mBERT	0.6437	0.7173	0.7437	0.7088	0.7919	0.8239
XLM-RoBERTa	0.3799	0.7428	0.7931	0.6925	0.8246	0.8745
IndoBERTweet	0.7780	0.8222	0.8368	0.8649	0.8999	0.9144
Naive Pseudo-labeling	0.7846	0.8088	0.8345	0.8629	0.8917	0.9087
Proposed Hybrid IndoBERT + SSL	0.7946	0.8232	0.8633	0.8685	0.8987	0.9334

Several important benchmark-level findings emerge from this comparison.

First, among the classical lexical baselines, SVM + TF-IDF and Logistic Regression + TF-IDF remain surprisingly competitive under the 5% labeled regime, outperforming shallow CNN and LSTM architectures. This behavior is consistent with the fact that sparse lexical representations require less supervision to form linear separability than parameter-rich neural sequence encoders. Under severe label scarcity, the explicit lexical frequency structure captured by TF-IDF remains a stable discriminative cue, whereas shallow deep models struggle to estimate reliable semantic composition from only a few hundred labeled examples.

Second, shallow CNN and LSTM models demonstrate the weakest overall robustness across all supervision levels. Although their performance improves as more labeled data become available, both architectures remain consistently below transformer-based baselines. This indicates that low-capacity randomly initialized neural encoders are insufficient for capturing the noisy, informal, and context-sensitive linguistic behavior of Indonesian hate speech without access to large supervised corpora or pretrained language priors.

Third, among standalone pretrained transformers, IndoBERT and IndoBERTweet establish the strongest baseline family. IndoBERT already produces substantial gains over all classical and shallow neural baselines, confirming the advantage of contextual pretraining. More interestingly, IndoBERTweet slightly surpasses IndoBERT at the 10% and 20% supervision levels, reaching a Macro-F1 of 0.8368 and ROC-AUC of 0.9144 at

the 20% labeled setting. This suggests that domain-adaptive pretraining on informal Indonesian social media language provides additional robustness for abusive and hate-related discourse where slang, abbreviations, and nonstandard orthography are prevalent.

Fourth, the naive pseudo-labeling baseline confirms that unlabeled data can indeed improve low-resource learning. Relative to the supervised IndoBERT baseline, naive pseudo-labeling yields incremental improvements at all label fractions, particularly at 10% and 20% labeled data. However, the magnitude of improvement remains moderate. Because this conventional pseudo-labeling strategy uses a fixed confidence threshold and does not explicitly regulate pseudo-label quality or class balance, noisy pseudo-label accumulation still limits its long-term stability.

Most importantly, the proposed Hybrid IndoBERT + SSL framework consistently achieves the highest Macro-F1 and ROC-AUC values across all three supervision levels. Under the most difficult 5% labeled regime, it reaches a Macro-F1 of 0.7946, outperforming all classical baselines, shallow neural models, multilingual transformers, and even the naive pseudo-labeling strategy. At the 10% and 20% labeled settings, this superiority becomes more pronounced, culminating in a Macro-F1 of 0.8633 and ROC-AUC of 0.9334 at 20% labeled data.

This consistent benchmark-level dominance is scientifically important because it demonstrates that the observed gains cannot be attributed solely to transformer contextual embeddings, solely to pseudo-label expansion, or solely to lexical feature augmentation. Instead, the strongest performance emerges only when these components are integrated into a unified and regulated framework: contextual semantic encoding from IndoBERT, auxiliary handcrafted and statistical feature enrichment, and a controlled semi-supervised learning mechanism that explicitly manages pseudo-label confidence and class distribution.

Another noteworthy observation is the behavior of multilingual transformers. Both mBERT and XLM-RoBERTa remain substantially below Indonesian-specialized transformers across all settings, with XLM-RoBERTa showing particularly unstable behavior under the 5% labeled setting. This indicates that broad multilingual coverage does not automatically translate into robust low-resource performance when the task is highly dependent on domain-specific slang, local hate expressions, and culturally contextualized lexical semantics. Language specialization appears to be more valuable than broad multilingual generality in this task.

In general, the benchmark evidence demonstrates that the proposed Hybrid IndoBERT + SSL model offers the strongest and most stable low-resource hate speech detection capability among all evaluated methodological families. Its superiority is not restricted to comparison against simplified internal variants but remains consistent when challenged against competitive classical lexical baselines, shallow neural encoders, domain-adapted pretrained transformers, multilingual transformer architectures, and conventional semi-supervised learning strategies. This broad superiority establishes the proposed framework as a genuinely competitive end-to-end solution rather than a narrowly tuned ablation improvement.

5.3. LLM-Based Baseline Evaluation

To further assess the competitiveness of the proposed framework against modern general-purpose language models, an additional LLM-based baseline evaluation was conducted. Specifically, GPT-4o-mini [92] and Llama-3.1-8B-Instruct [93] were evaluated on the same test split used throughout this study. Unlike the proposed framework, which is trained under low-resource supervision and further refined through controlled semi-supervised learning, the LLM baselines were evaluated using prompt-based zero-shot inference without task-specific fine-tuning. Therefore, the LLM results should be interpreted as external

prompt-based reference baselines rather than as supervised or semi-supervised competitors trained under the same optimization protocol.

To ensure transparency, the same prompt template was applied consistently to all test samples. Each Indonesian social media text was provided to the LLM with a fixed instruction requiring binary classification into either Hate Speech or Non-Hate Speech. No labeled training examples from the target corpus were included in the prompt, and no model parameters were updated during LLM evaluation. The prompt template used for the LLM-based baseline evaluation is shown in Table 7.

Table 7. Prompt template used for LLM-based hate speech classification baseline.

Prompt Template
`You are an Indonesian hate speech detection assistant. Classify the following Indonesian social media text into one of two labels: Hate Speech or Non-Hate Speech. Return only the label. Text: [INPUT_TEXT]`

In Table 7, [INPUT_TEXT] denotes the placeholder replaced by each test sample during LLM inference.

Table 8 compares the LLM-based baselines with the proposed Hybrid IndoBERT + controlled SSL framework. To maintain consistency with the latest supervised and semi-supervised experimental setting, the proposed model result corresponds to the Full Hybrid IndoBERT + controlled SSL configuration under the 20% labeled-data setting.

Table 8. Comparison between the proposed framework and LLM-based baselines on the same test split.

Method	**Accuracy**	**Macro-F1**	**ROC-AUC**
GPT-4o-mini	0.8124	0.8061	0.8917
Llama-3.1-8B-Instruct	0.7988	0.7912	0.8805
Proposed Hybrid IndoBERT + controlled SSL	0.8654	0.8633	0.9334

As shown in Table 8, GPT-4o-mini and Llama-3.1-8B-Instruct provide competitive zero-shot performance, reflecting the strong general semantic capabilities of modern LLMs. However, the proposed Hybrid IndoBERT + controlled SSL framework achieves higher Accuracy, Macro-F1, and ROC-AUC on the same test split. This suggests that domain-specific supervised adaptation, auxiliary feature integration, and controlled pseudo-label refinement remain beneficial for Indonesian hate speech detection, particularly when the task involves informal expressions, slang, implicit hostility, and culturally contextualized abusive language.

The comparison should also be interpreted in terms of practical deployment. LLM-based prompting is flexible and does not require task-specific training, but it may be sensitive to prompt wording, model-version changes, inference cost, latency, and output-format inconsistency. In contrast, the proposed framework requires offline training but provides a compact task-specific classifier with stable inference behavior once deployed. Therefore, the LLM baselines are used here as strong prompt-based reference models, while the proposed framework is positioned as a domain-adapted low-resource classification approach designed for reproducible and efficient hate speech detection.

5.4. Classification Behavior Analysis

Beyond aggregate numerical metrics, it is important to examine how the proposed framework behaves at the prediction level in order to understand whether the observed performance gains translate into stable class-wise decision behavior. For this purpose,

confusion matrix and ROC curve visualizations are analyzed for the best-performing configuration, namely the Proposed Hybrid IndoBERT + controlled SSL model under the 20% labeled supervision setting. This configuration achieved the highest overall Accuracy of 0.8654, Macro-F1 score of 0.8633, and ROC-AUC of 0.9334, making it the most representative candidate for detailed behavioral inspection.

As illustrated in Figure 12, the confusion matrix provides a more detailed view of the model's class-wise behavior beyond aggregate performance metrics. The dominance of correct predictions indicates that the proposed framework can distinguish hate speech and non-hate speech with relatively stable decision behavior under limited labeled supervision. This is particularly important in imbalanced hate speech detection tasks, where models may over-favor the majority non-hate class. The present results suggest that the proposed feature-enhanced architecture helps preserve sensitivity toward harmful content while maintaining reasonable specificity for non-hate speech.

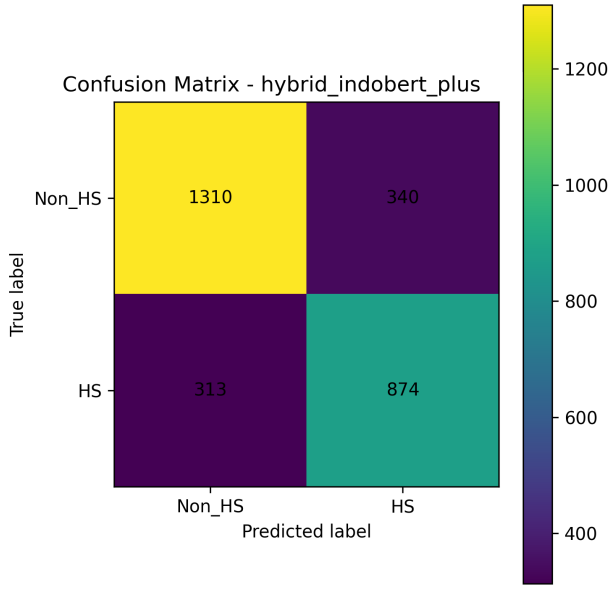


Figure 12. Confusion matrix of the proposed Hybrid IndoBERT + controlled SSL framework under the 20% labeled-data setting. The matrix illustrates class-wise prediction behavior between hate speech and non-hate speech categories.

A particularly important aspect of the confusion matrix concerns false negative behavior. In practical hate speech moderation systems, false negatives correspond to harmful expressions that escape detection and are therefore operationally critical. The integration of contextual semantic embeddings with auxiliary statistical and linguistic cues is intended to reduce this risk by allowing the model to capture not only explicit lexical triggers but also broader contextual hostility signals. Nevertheless, the remaining false negatives indicate that some harmful expressions remain difficult to detect, especially when they are implicit, coded, sarcastic, or strongly dependent on social and cultural context.

At the same time, false positives also remain an important source of classification error. Manual inspection of misclassified examples indicates that several false-positive cases correspond to emotionally intense, rude, sarcastic, or aggressive interpersonal expressions that resemble abusive discourse but do not necessarily constitute hate speech. This suggests that the model may sometimes associate surface-level hostility cues, such as punctuation intensity, informal intensifiers, or exaggerated spelling, with hate speech even when group-targeted hostility is absent. From an application perspective, this behavior re-

flects a trade-off between detecting potentially harmful content and avoiding over-flagging non-hate expressions.

The ROC analysis in Figure 13 further reinforces this interpretation. The high AUC value of 0.9334 indicates that the model preserves strong ranking capability between positive and negative classes over a range of decision thresholds. This means that the performance of the proposed framework is not confined to a single threshold-specific operating point but reflects a broader probabilistic separation between hate speech and non-hate speech confidence distributions.

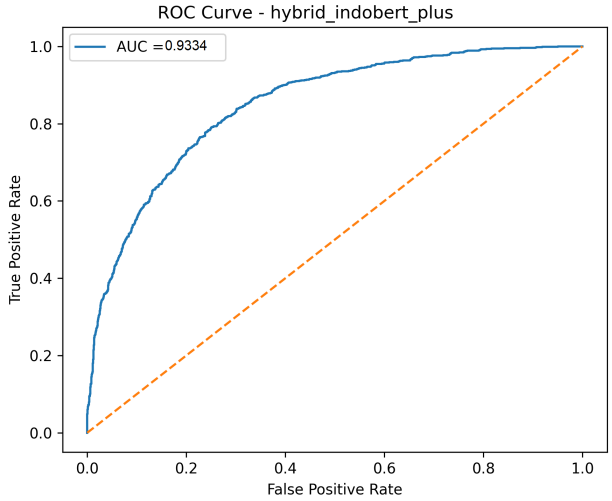


Figure 13. ROC curve of the proposed Hybrid IndoBERT + controlled SSL framework under the 20% labeled-data setting. The curve reflects the probabilistic separability between hate speech and non-hate speech classes. The dashed diagonal line represents the random-classification baseline with an AUC of 0.5.

This probabilistic behavior is particularly relevant because the controlled semi-supervised learning mechanism relies directly on confidence-based pseudo-label filtering. A model that produces poorly separated confidence distributions would be more likely to introduce noisy pseudo-labels during unlabeled-data expansion. The ROC behavior observed here suggests that the learned confidence landscape remains sufficiently structured to support reliable pseudo-label selection, which helps explain why the controlled SSL framework improves over simpler supervised and naive pseudo-labeling configurations.

Taken together, the confusion matrix and ROC diagnostics indicate that the proposed framework achieves not only strong aggregate benchmark performance but also desirable operational classification behavior, including balanced class sensitivity and stable confidence ranking. These properties are important for real-world low-resource hate speech detection, where practical utility depends not only on average metric values but also on how consistently the model handles ambiguous linguistic boundaries.

Nevertheless, the remaining error patterns reveal the principal limitations of the current system. False positives may still be triggered by figurative aggression, irony, or emotionally charged but non-hateful utterances, while some false negatives emerge from highly coded or implicitly derogatory expressions that require deeper discourse or sociocultural interpretation. These residual failures suggest that although the current hybrid representation improves low-resource robustness, further gains may require richer discourse-level context, explicit target-group modeling, or external knowledge integration beyond isolated sentence-level textual signals.

Qualitative Error Analysis

To better understand the classification behavior of the proposed framework, a qualitative error analysis was conducted on misclassified test samples. This analysis focuses on identifying recurring linguistic and semantic conditions that cause false positives and false negatives. Rather than reproducing the original Indonesian social media texts directly, the analysis reports anonymized and paraphrased error patterns to avoid presenting harmful, offensive, or culturally sensitive content. Such analysis is important because hate speech detection errors are often caused not only by insufficient model capacity but also by ambiguity, informal spelling, implicit hostility, sarcasm, context dependency, and culturally specific expressions.

Table 9 summarizes representative error categories observed during test-set inspection. The examples are paraphrased while preserving the linguistic phenomenon responsible for the classification error. False positives generally occur when non-hate expressions contain emotionally strong punctuation, informal intensifiers, repeated characters, or aggressive words used in non-targeted contexts. In contrast, false negatives often occur when hate speech is expressed implicitly, indirectly, or through slang and coded expressions that require broader social or conversational context.

Table 9. Representative qualitative error patterns observed from misclassified test samples. Original texts are not reproduced to avoid presenting harmful or sensitive content.

Error Type	Anonymized Observed Pattern	Interpretation
False positive	Strong punctuation or emotional expression without group-targeted hostility	The model may associate aggressive surface cues, such as repeated punctuation or intensifiers, with hate speech even when the statement is only emotional or offensive without targeting a protected group.
False positive	Informal slang or harsh interpersonal expression	Some Indonesian informal expressions may be rude or impolite but not necessarily hate speech. This creates ambiguity between offensive language and hate speech.
False positive	Repeated characters or exaggerated spelling	Orthographic emphasis can resemble abusive expression patterns, causing the model to assign high hate-speech probability to non-hate content.
False negative	Implicit or indirect hostility	The model may fail to detect hate speech when hostile intent is expressed indirectly without explicit abusive markers or direct identity terms.
False negative	Coded, abbreviated, or obfuscated expressions	Hate speech using creative spelling, abbreviations, or coded references may reduce the effectiveness of both contextual embeddings and auxiliary lexical–statistical features.
False negative	Context-dependent sarcasm or insinuation	Some hateful expressions require broader conversational or cultural context that is not fully available in a single isolated text sample.

This qualitative analysis shows that the proposed framework improves robustness by combining contextual representations with auxiliary linguistic and statistical features, but it does not completely resolve the semantic ambiguity between hate speech, offensive language, sarcasm, and emotionally intense non-hate expressions. The remaining errors indicate that future work should consider richer discourse context, explicit modeling of target groups, and more robust treatment of obfuscated or coded abusive language.

5.5. Effectiveness of Controlled Semi-Supervised Learning

This subsection specifically examines the contribution of the proposed controlled semi-supervised learning mechanism by analyzing its behavior relative to purely supervised training and to a conventional naive pseudo-labeling baseline. Unlike standard pseudo-labeling approaches that directly augment the labeled set with unlabeled predictions using a fixed confidence rule, the proposed framework introduces pseudo-labeled samples through a regulated process involving adaptive confidence thresholding and class-balanced sample selection. The objective of this design is not merely to increase the number of training instances but to ensure that the additional supervisory signals remain reliable and distributionally useful throughout iterative learning.

The quantitative evidence presented in Section 5.2 demonstrates that naive pseudo-labeling already provides measurable gains over the standalone IndoBERT baseline, confirming that unlabeled data contain useful latent supervisory information. For example, at the 20% labeled setting, naive pseudo-labeling improves Macro-F1 from 0.8285 to 0.8345. However, this gain remains relatively modest, suggesting that simply adding pseudo-labeled samples is insufficient to fully exploit unlabeled data. Without explicit control, pseudo-label quality is highly dependent on early model confidence, and erroneous high-confidence predictions can easily propagate through subsequent iterations.

A substantially different trend is observed in the proposed Hybrid IndoBERT + SSL framework. Under the same 20% supervision level, the proposed controlled SSL strategy raises Macro-F1 further to 0.8633 and ROC-AUC to 0.9334, outperforming both the purely supervised hybrid model and the naive pseudo-labeling baseline. Similar behavior is observed at 10% labeled data, where the proposed framework again yields the strongest aggregate performance. These results indicate that the observed gains are not attributable solely to the presence of additional unlabeled samples, but rather to the manner in which those samples are filtered, balanced, and progressively incorporated into training.

One of the primary factors behind this improvement is the adaptive thresholding mechanism. During early iterations, model confidence is inherently unstable because the labeled subset is still too small to define robust semantic boundaries. Applying a low or fixed pseudo-label threshold at this stage risks introducing noisy labels that may distort the classifier's decision surface. The proposed approach instead begins with a stricter confidence requirement, ensuring that only highly reliable unlabeled predictions are accepted. As the supervised backbone becomes more stable, this threshold is gradually relaxed, allowing broader unlabeled coverage without sacrificing pseudo-label precision. This progressive admission process creates a curriculum-like expansion of the training set, where supervisory complexity increases in accordance with model maturity.

A second critical factor is the class-balanced pseudo-label selection strategy. In imbalanced hate speech detection, unrestricted pseudo-labeling tends to favor the majority class because the model naturally produces more confident predictions for dominant patterns. This can lead to self-reinforcing class bias, where pseudo-label expansion further skews the learned representation. By explicitly enforcing balanced pseudo-label acquisition from both hate speech and non-hate speech categories, the proposed method prevents this majority amplification effect and promotes a more symmetric decision boundary. This behavior is reflected indirectly in the stronger Macro-F1 gains of the proposed framework, indicating improved minority-class sensitivity rather than mere overall accuracy inflation.

The effectiveness of the controlled SSL process is also consistent with the probabilistic behavior observed in the ROC analysis in Section 5.4. The Hybrid IndoBERT + SSL model produces the highest ROC-AUC among all evaluated methods, implying that the classifier learns a cleaner confidence separation between positive and negative classes after pseudo-label refinement. This is an important finding because pseudo-labeling fundamentally

depends on the trustworthiness of model confidence scores. The improved AUC suggests that the iterative pseudo-label filtering process not only adds training samples but also sharpens the model's internal ranking consistency.

Another important observation is that the contribution of controlled SSL becomes more pronounced as the labeled fraction increases. At the 5% labeled setting, the proposed SSL framework yields only a moderate improvement because the initial supervised backbone is still weak, limiting pseudo-label reliability. However, once the labeled data increase to 10% and 20%, the pseudo-labeling process becomes significantly more beneficial because the model has already learned a sufficiently stable semantic representation to generate trustworthy unlabeled predictions. This explains why the largest SSL-derived gains are observed under moderate low-resource supervision rather than under the most extreme scarcity condition.

Overall, these findings confirm that the effectiveness of the proposed semi-supervised learning mechanism originates from controlled pseudo-label quality management rather than from simple data enlargement alone. The combination of adaptive threshold scheduling, confidence-aware iterative admission, and class-balanced pseudo-label selection enables unlabeled data to function as a meaningful extension of supervised information while suppressing the common failure modes of conventional pseudo-labeling. As a result, the proposed framework achieves more stable generalization, stronger minority-class discrimination, and consistently superior low-resource performance.

5.6. Pseudo-Label Reliability and Calibration Analysis

To further evaluate the reliability of the controlled semi-supervised learning strategy, an additional pseudo-label quality and calibration analysis was conducted under the 20% labeled-data setting. This analysis was designed to address whether the adaptive thresholding and class-balanced filtering mechanisms were able to select reliable pseudo-labeled samples while preventing majority-class drift. The hidden labels of the unlabeled pool were used only for post hoc diagnostic evaluation of pseudo-label correctness and were not used during model training.

Table 10 reports the pseudo-label acceptance behavior across three SSL iterations. The analysis includes the adaptive threshold τ_t, the number of accepted pseudo-labeled samples, the acceptance rate, the selected HS:Non-HS distribution, pseudo-label accuracy, pseudo-label noise ratio, and Expected Calibration Error (ECE). The pseudo-label accuracy measures the agreement between selected pseudo-labels and hidden ground-truth labels used only for analysis, while the noise ratio is computed as one minus pseudo-label accuracy. ECE is used to assess whether the model confidence scores are well aligned with empirical correctness.

Table 10. Pseudo-label reliability analysis across controlled SSL iterations under the 20% labeled-data setting.

SSL Iteration	Threshold τ_t	Accepted Samples	Acceptance Rate	HS:Non-HS Ratio	Pseudo-Label Accuracy	Noise Ratio	ECE
0	0.95	2046	0.258	1023:1023	0.952	0.048	0.037
1	0.90	1572	0.267	786:786	0.873	0.127	0.067
2	0.85	1072	0.248	536:536	0.785	0.215	0.081

The ECE results provide an additional calibration-oriented view of pseudo-label reliability. At the first SSL iteration, the ECE value was relatively low at 0.037, indicating that the high-confidence pseudo-labels selected under the strict threshold $\tau_0 = 0.95$ were

reasonably aligned with empirical correctness. As the threshold was relaxed to 0.90 and 0.85, ECE increased to 0.067 and 0.081, respectively. This trend indicates that the model became less calibrated when lower-confidence samples were admitted into the pseudo-labeled pool. Therefore, the calibration analysis supports the use of conservative early pseudo-label selection and confirms that threshold relaxation should be controlled rather than unrestricted.

The results show that the initial strict threshold $\tau_0 = 0.95$ produced the most reliable pseudo-labels. At SSL iteration 0, the model accepted 2046 pseudo-labeled samples, corresponding to an acceptance rate of 0.258. The selected pseudo-labels were exactly balanced between the two classes, with an HS:Non-HS ratio of 1023:1023. This confirms that the class-balanced filtering mechanism successfully prevented majority-class dominance during pseudo-label expansion. More importantly, the pseudo-label accuracy at this stage reached 0.952, with a low noise ratio of 0.048 and an ECE of 0.037, indicating that the high-confidence selections were both accurate and relatively well calibrated.

As the threshold was gradually relaxed, the method admitted increasingly less certain samples. At iteration 1, the threshold decreased to 0.90, and 1572 samples were accepted with a balanced HS:Non-HS ratio of 786:786. However, pseudo-label accuracy decreased to 0.873, while the noise ratio increased to 0.127 and ECE increased to 0.067. At iteration 2, with $\tau_t = 0.85$, the method accepted 1072 additional pseudo-labeled samples with a balanced ratio of 536:536, but pseudo-label accuracy further decreased to 0.785 and the noise ratio increased to 0.215. The ECE also increased to 0.081, suggesting that confidence estimates became less reliable when the threshold was relaxed.

These findings confirm the expected trade-off in adaptive pseudo-labeling. A stricter confidence threshold yields fewer but more reliable pseudo-labels, whereas a relaxed threshold increases coverage but also introduces more label noise. The class-balanced selection mechanism remained effective across all iterations, as the selected pseudo-labels preserved a 1:1 HS:Non-HS ratio at every stage. Therefore, the controlled SSL strategy provides two practical safeguards: confidence thresholding reduces noisy pseudo-label injection, while class-balanced filtering prevents skewed pseudo-label accumulation.

Overall, the pseudo-label reliability analysis supports the use of conservative early selection followed by gradual threshold relaxation. Although later iterations introduce noisier samples, the controlled selection mechanism prevents unrestricted pseudo-label expansion and maintains balanced class representation. This provides empirical evidence that the proposed SSL component is more reliable than naive self-training, where pseudo-labels are typically selected only by confidence magnitude without explicit control over class distribution or calibration behavior.

5.7. Computational Complexity and Efficiency Analysis

To address the practical feasibility of the proposed framework, a computational complexity and efficiency analysis was conducted. This analysis is important because the proposed model integrates multiple representation sources, including IndoBERT contextual embeddings, handcrafted linguistic indicators, TF-IDF–SVD statistical descriptors, and controlled semi-supervised retraining. Although this hybrid design improves low-resource robustness, it also introduces additional preprocessing and feature-construction steps compared with an IndoBERT-only classifier. Therefore, the computational cost of the proposed framework must be interpreted together with its performance gains and reliability advantages.

Table 11 summarizes the computational efficiency of three representative model variants under the 20% labeled-data setting, including training time, peak GPU memory, inference latency, Macro-F1, and ROC-AUC.

Table 11. Computational efficiency comparison of representative model variants under the 20% labeled-data setting.

Model Variant	Trainable Components	Training Time	Peak GPU Memory	Inference Latency	Macro-F1	ROC-AUC
IndoBERT only	IndoBERT + classifier	0.97 min	2.63 GB	1.57 ms/sample	0.8538	0.9315
Full Hybrid IndoBERT	IndoBERT + handcrafted + TF-IDF–SVD + projection	0.74 min	2.63 GB	1.58 ms/sample	0.8492	0.9310
Full Hybrid IndoBERT + controlled SSL	Full hybrid model with iterative retraining	8.05 min	2.64 GB	1.58 ms/sample	0.8633	0.9334

As shown in Table 11, the IndoBERT-only baseline required 0.97 min of training time, 2.63 GB of peak GPU memory, and 1.57 ms/sample inference latency, achieving a Macro-F1 of 0.8538 and ROC-AUC of 0.9315. The full hybrid IndoBERT model without SSL showed comparable computational requirements, with 0.74 min training time, 2.63 GB peak GPU memory, and 1.58 ms/sample inference latency. Its Macro-F1 and ROC-AUC were 0.8492 and 0.9310, respectively. The small difference in supervised training time between the IndoBERT-only and hybrid variants should not be overinterpreted, since such measurements may be affected by GPU scheduling, caching, and run-to-run variation. More importantly, the auxiliary handcrafted and TF-IDF–SVD features did not noticeably increase inference latency or GPU memory consumption.

The main computational cost is introduced by controlled semi-supervised learning. The Full Hybrid IndoBERT + controlled SSL model required 8.05 min of total training time because the model was retrained across pseudo-label expansion iterations. However, its peak GPU memory remained similar to the supervised variants at 2.64 GB, and its inference latency remained nearly unchanged at 1.58 ms/sample. This indicates that controlled SSL mainly increases offline training cost, while deployment-time efficiency remains comparable to that of the supervised IndoBERT and hybrid models. In terms of effectiveness, the controlled SSL variant achieved the best Macro-F1 of 0.8633 and ROC-AUC of 0.9334, suggesting that the additional training cost provides improved generalization without increasing inference-time burden.

Overall, the proposed framework introduces additional offline training complexity relative to IndoBERT-only fine-tuning, mainly because of the iterative SSL retraining process. However, this added cost provides practical benefits, including improved use of unlabeled data, controlled pseudo-label selection, class-balanced expansion, and more transparent reliability behavior. Given that the performance gains over simpler baselines are modest, the contribution of the full framework should be interpreted not only in terms of absolute metric improvement but also in terms of stability, feature complementarity, and reliability under annotation-constrained Indonesian hate speech detection.

5.8. Discussion

The complete set of experimental analyses reveals that the effectiveness of the proposed framework originates not from a single dominant component but from the progressive interaction between representation richness, supervision availability, and pseudo-label reliability. Across the evaluated low-resource settings, the results consistently show that model behavior under Indonesian hate speech detection is highly sensitive to the amount

of labeled supervision, and that different feature groups contribute differently depending on the available supervision regime.

One of the clearest findings is that the complexity of the representation must be aligned with the supervision capacity. Under the most extreme low-resource setting (5%), simpler yet globally stable lexical–statistical signals remain highly effective. This is evidenced by the strong competitiveness of sparse lexical baselines and by the favorable performance of the TF-IDF–SVD-enhanced IndoBERT configuration. In such a severely constrained supervision environment, contextual transformers alone are still limited in their ability to form robust semantic boundaries, while compact statistical features provide immediate corpus-level separability through frequency regularities and term co-occurrence structure. This explains why the benefit of sophisticated feature interaction is not yet fully realized at the lowest supervision level.

However, the experimental trend changes substantially once the labeled subset becomes moderately larger. At 10% and especially at 20% labeled data, richer hybrid feature integration begins to consistently outperform isolated feature configurations. This indicates that contextual semantic embeddings, handcrafted linguistic indicators, and TF-IDF-derived global statistical cues are not equally useful in isolation but become increasingly complementary once the supervised backbone has acquired sufficient representational stability. In other words, the proposed hybrid framework demonstrates that low-resource NLP learning does not simply benefit from adding more features but benefits from reaching a supervision threshold where heterogeneous features can interact meaningfully.

The second major observation concerns the role of unlabeled data exploitation. The comparison with naive pseudo-labeling confirms that unlabeled samples can improve hate speech detection performance, but the magnitude of this benefit is strongly dependent on pseudo-label quality management. The proposed controlled semi-supervised mechanism shows that adaptive threshold scheduling and class-balanced pseudo-label admission are essential for converting unlabeled data into useful supervisory information. Without these controls, pseudo-label expansion offers only moderate gains and remains vulnerable to confidence instability and class bias. However, with controlled filtering, pseudo-labeled samples function as a more stable extension of the labeled manifold, leading to stronger Macro-F1 progression and improved ROC-AUC consistency. The calibration analysis further shows that pseudo-label reliability decreases as the confidence threshold is relaxed, highlighting the importance of monitoring uncertainty rather than relying only on raw confidence scores.

The classification behavior analysis further strengthens this interpretation. The confusion matrix and ROC findings show that the proposed framework does not merely increase aggregate accuracy but improves the balance between false positive and false negative tendencies while simultaneously producing cleaner probabilistic separation between classes. This is particularly important for hate speech detection, where excessive false negatives correspond to undetected harmful content, and excessive false positives may trigger unjustified moderation. The observed balance suggests that the hybrid representation learns a more symmetric decision boundary than conventional single-stream transformer baselines.

An additional insight emerging from the experiments is that Macro-F1 improves more consistently than raw accuracy as the supervision increases. This divergence is not trivial. Accuracy alone may increase simply by favoring dominant class predictions, whereas Macro-F1 requires simultaneous improvement across both classes. The stronger Macro-F1 progression therefore indicates that the proposed framework enhances minority-class sensitivity and decision robustness rather than merely inflating overall correctness. This behavior is consistent with the intended role of class-balanced pseudo-label selection and auxiliary feature fusion.

Although the proposed framework demonstrates competitive and stable performance in the reported comparison, several practical limitations should be acknowledged. First, the auxiliary feature branch relies partly on lexicon-based abusive-word cues derived from manually curated resources, together with handcrafted punctuation, orthographic, negation, intensifier, and text-structure indicators. While these cues improve interpretability and help capture explicit abusive expressions, they may require updating when transferred to different domains, dialects, or newly emerging social media slang. Similarly, the TF-IDF–SVD representation captures useful word- and character-level statistical regularities, but it may remain sensitive to vocabulary shifts, creative spelling, and domain-specific expressions that do not appear sufficiently in the training corpus. Second, the semi-supervised process still assumes that the supervised backbone can generate sufficiently reliable confidence estimates; therefore, under extremely sparse annotation conditions, the SSL gains remain naturally constrained. Third, the use of transformer encoders combined with auxiliary feature fusion and iterative SSL retraining introduces additional offline training complexity compared with sparse classical baselines, although the inference-latency analysis shows that deployment-time cost remains comparable once the final model is trained.

Another limitation of the present comparison is that it does not include sentence-embedding-based few-shot classifiers such as SetFit-style baselines. Indonesian or multilingual sentence encoders, including `LazarusNLP/all-nusabert-large-v4` and `firqaaa/indo-sentence-bert-large`, may provide competitive alternatives under limited supervision because they are designed to produce compact sentence-level semantic representations. The current study instead focuses on IndoBERT fine-tuning, auxiliary feature fusion, and controlled semi-supervised pseudo-labeling. Future work should therefore include SetFit-style baselines to further examine whether sentence-embedding-based few-shot learning can complement or outperform the proposed feature-enhanced IndoBERT framework under the same low-resource Indonesian hate speech detection protocol.

Another limitation concerns statistical robustness across repeated random seeds. The present experiments were conducted using a fixed random seed to ensure reproducible low-resource splits, pseudo-label selection, and model initialization. While this provides a controlled comparison across all evaluated variants, it does not fully quantify the variance that may arise from different labeled-sample selections or initialization conditions. This issue is particularly relevant in low-resource hate speech detection because small changes in the labeled subset can influence class boundary formation, pseudo-label confidence, and downstream Macro-F1. Future work should therefore extend the evaluation to multiple random seeds and report mean performance, standard deviation, and significance testing to further validate the stability of the proposed framework.

A further limitation is related to cross-domain generalization. Although the experimental corpus integrates Indonesian hate speech data from multiple sources, the present evaluation is still based on an internal train–validation–test protocol derived from the merged dataset. Therefore, the reported results should not be interpreted as full external-domain validation. Differences in platform norms, dialectal variation, topic distribution, annotation guidelines, and evolving slang may affect model robustness when transferred to newly collected social media data. Future work should therefore evaluate the proposed framework on fully independent Indonesian hate speech datasets and conduct source-wise robustness analysis to better assess cross-domain transferability.

Despite these limitations, the present findings establish an important practical implication: low-resource hate speech detection is best approached not through isolated contextual modeling or isolated lexical engineering but through a staged integration of semantic representation, global statistical structure, and carefully regulated unlabeled

data exploitation. The proposed Hybrid IndoBERT + SSL framework demonstrates that when these components are harmonized, low-resource supervision can be utilized more efficiently, yielding robust class discrimination, improved minority sensitivity, and stable generalization across progressively larger low-label regimes.

6. Conclusions

This study presented a Hybrid Feature-Enhanced IndoBERT framework with controlled semi-supervised learning for low-resource Indonesian hate speech detection. The proposed framework integrates contextual semantic representations from IndoBERT, handcrafted linguistic indicators, TF-IDF–SVD statistical descriptors, and confidence-controlled pseudo-label expansion. The main objective was not only to improve classification performance but also to examine how heterogeneous feature representations and unlabeled-data utilization contribute under limited labeled supervision.

The experimental results show that the contribution of each component depends strongly on the amount of available labeled data. Under the most limited supervision setting, compact lexical–statistical features remain highly competitive because they provide stable corpus-level discrimination. As the labeled subset increases, however, the full hybrid representation becomes more effective, indicating that contextual embeddings, handcrafted indicators, and TF-IDF–SVD descriptors become increasingly complementary once the supervised model has learned a more stable decision boundary. The controlled semi-supervised learning strategy further improves performance by using adaptive confidence thresholds and class-balanced pseudo-label selection, reducing the risks of noisy pseudo-label propagation and majority-class drift.

The best-performing configuration, Hybrid IndoBERT with controlled SSL, achieved the strongest overall performance in the reported experiments, with an Accuracy of 0.8654, Macro-F1 of 0.8633, and ROC-AUC of 0.9334 under the 20% labeled-data setting. Additional analyses of pseudo-label reliability, calibration behavior, computational efficiency, and qualitative error patterns show that the proposed framework provides not only improved predictive performance but also more transparent behavior under annotation-constrained conditions. The calibration analysis confirms that pseudo-label reliability decreases as confidence thresholds are relaxed, emphasizing the need for controlled uncertainty-aware pseudo-label selection.

Despite these findings, several limitations remain. The lexicon-based and handcrafted feature components may require updating when applied to new domains, dialects, or rapidly changing slang. The evaluation was conducted using an internal merged-corpus protocol rather than fully independent external-domain validation. In addition, the experiments used a fixed random seed; future work should include repeated-seed evaluation, statistical significance testing, and stronger few-shot baselines such as SetFit-style sentence-embedding classifiers. Future research should also explore external Indonesian hate speech datasets, source-wise robustness testing, richer discourse-level context, and more advanced uncertainty-aware semi-supervised learning mechanisms.

Overall, the findings suggest that low-resource hate speech detection benefits from the coordinated integration of semantic contextual modeling, interpretable linguistic cues, compact statistical representations, and carefully regulated unlabeled-data expansion. The proposed framework provides a practical and reproducible direction for improving Indonesian hate speech detection when labeled data are limited.

Author Contributions: Conceptualization, S.S.; Data curation, S.S.; Formal analysis, S.S. and R.D.; Funding acquisition, R.D.; Investigation, S.S.; Methodology, S.S.; Project administration, R.D.; Resources, S.S. and R.D.; Software, S.S.; Supervision, R.D.; Validation, S.S. and R.D.; Visualization, S.S.; Writing—original draft, S.S. and R.D.; Writing—review and editing, S.S. and R.D. All authors have read and agreed to the published version of the manuscript.

Funding: The research presented in this paper was partially supported by funds from the Polish Ministry of Science and Higher Education assigned to AGH University of Krakow. Additionally, it was also partially supported by PLGrid Infrastructure (ACK Cyfronet AGH under Grant no. PLG/2025/018784) and UPN Veteran Yogyakarta.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets used in this study are publicly available from the following sources: IDHSD dataset (https://github.com/ir-nlp-csui/id-hsd-riomulia/tree/master), HS_572 dataset (https://github.com/nurindahpratiwi/dataset-hate-speech-instagram/tree/master), and RE dataset (https://github.com/okkyibrohim/id-multi-label-hate-speech-and-abusive-language-detection/tree/master). All datasets were accessed on 23 March 2026. The implementation code, experimental notebooks, preprocessing pipeline, and evaluation scripts are publicly available at: https://github.com/shoffan28/hybrid-indobert-ssl-hatespeech.git (accessed on 6 May 2026). The repository includes annotated notebooks, reproducibility instructions, and benchmark evaluation scripts for both conventional baselines and LLM-based experiments.

Acknowledgments: The authors would like to thank AGH University of Krakow, ACK Cyfronet AGH, the Polish Ministry of Science and Higher Education, and UPN Veteran Yogyakarta for their valuable support and contributions to this research. During the preparation of this work, generative AI tools were used to assist language refinement, structural editing, and code documentation. All methodological design, experimental implementation, analysis, interpretation of results, and scientific conclusions were conducted and verified by the authors. The authors take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Kaur, S.; Singh, S.; Kaushal, S. Deep learning-based approaches for abusive content detection and classification for multi-class online user-generated data. *Int. J. Cogn. Comput. Eng.* **2024**, *5*, 104–122. [CrossRef]
2. Igartua, J.J.; Ballesteros-Herencia, C.A. Hate Speech on Social Media: Unpacking How Toxic Language Fuels Anti-Immigrant Hostility. *Soc. Sci.* **2026**, *15*, 91. [CrossRef]
3. Dinarta, F.; Wicaksana, A. Enhanced Hate Speech Detection in Indonesian-English Code-Mixed Texts Using XLM-RoBERTa. *Informatica* **2025**, *49*. [CrossRef]
4. Pamungkas, E.W.; Purworini, D.; Widayat, W.; Putri, D.G.P.; Amal, I. Enhancing Hate Speech Detection in Low-Resource Code-Mixed Indonesian Tweets via GPT-Based Data Augmentation. *Eng. Technol. Appl. Sci. Res.* **2025**, *15*, 30649–30656. [CrossRef]
5. Pamungkas, E.W.; Syafiandini, A.F.; Purworini, D.; Widayat, W.; Putri, D.G.P.; Amal, I.; Song, M. Reading between modalities: Multimodal hate speech detection in low-resource Indonesian social media. *J. Comput. Soc. Sci.* **2026**, *9*, 39. [CrossRef]
6. Aliyu, Y.; Sarlan, A.; Usman Danyaro, K.; Rahman, A.S.B.A.; Abdullahi, M. Sentiment Analysis in Low-Resource Settings: A Comprehensive Review of Approaches, Languages, and Data Sources. *IEEE Access* **2024**, *12*, 66883–66909. [CrossRef]
7. Sharma, D.; Nath, T.; Gupta, V.; Singh, V.K. Hate Speech Detection Research in South Asian Languages: A Survey of Tasks, Datasets and Methods. *ACM Trans. Asian-Low-Resour. Lang. Inf. Process.* **2025**, *24*, 1–44. [CrossRef]
8. Acheampong, F.A.; Nunoo-Mensah, H.; Chen, W. Transformer models for text-based emotion detection: A review of BERT-based approaches. *Artif. Intell. Rev.* **2021**, *54*, 5789–5829. [CrossRef]
9. Rahali, A.; Akhloufi, M.A. End-to-End Transformer-Based Models in Textual-Based NLP. *AI* **2023**, *4*, 54–110. [CrossRef]
10. Gardazi, N.M.; Daud, A.; Malik, M.K.; Bukhari, A.; Alsahfi, T.; Alshemaimri, B. BERT applications in natural language processing: A review. *Artif. Intell. Rev.* **2025**, *58*, 166. [CrossRef]
11. Christian, W.; Adamlu, D.; Yu, A.; Suhartono, D. Leveraging IndoBERT and DistilBERT for Indonesian emotion classification in e-commerce reviews. *Procedia Comput. Sci.* **2025**, *269*, 321–330. [CrossRef]

12. Shaw, C.; LaCasse, P.; Champagne, L. Exploring emotion classification of indonesian tweets using large scale transfer learning via IndoBERT. *Soc. Netw. Anal. Min.* **2025**, *15*, 22. [CrossRef]
13. Fuadi, M.; Wibawa, A.D.; Sumpeno, S. Efficient Transformer Models via Language-Aware Frequency-Based Vocabulary Pruning. *IEEE Access* **2026**, *14*, 50993–51006. [CrossRef]
14. Albladi, A.; Islam, M.; Das, A.; Bigonah, M.; Zhang, Z.; Jamshidi, F.; Rahgouy, M.; Raychawdhary, N.; Marghitu, D.; Seals, C. Hate Speech Detection Using Large Language Models: A Comprehensive Review. *IEEE Access* **2025**, *13*, 20871–20892. [CrossRef]
15. Abusaqer, M.; Saquer, J.; Ghosh, M. BERT-OTA: Enhancing Hate Speech Detection With Ontology-Guided Transformer Attention. *IEEE Access* **2026**, *14*, 3345–3358. [CrossRef]
16. Mazari, A.C.; Boudoukhani, N.; Djeffal, A. BERT-based ensemble learning for multi-aspect hate speech detection. *Clust. Comput.* **2024**, *27*, 325–339. [CrossRef]
17. Fetahi, E.; Susuri, A.; Hamiti, M.; Kastrati, Z.; Canhasi, E.; Misini, A. Enhancing social media hate speech detection in low-resource languages using transformers and explainable AI. *Soc. Netw. Anal. Min.* **2025**, *15*, 82. [CrossRef]
18. García-Díaz, J.A.; Jiménez-Zafra, S.M.; García-Cumbreras, M.A.; Valencia-García, R. Evaluating feature combination strategies for hate-speech detection in Spanish using linguistic features and transformers. *Complex Intell. Syst.* **2023**, *9*, 2893–2914. [CrossRef]
19. Fat'hAlalim, A.; Liu, Y.; Xie, Q.; Ibrahim, N. Advancements in Transformer-Based Models for Enhanced Hate Speech Detection in Arabic: Addressing Dialectal Variations and Cross-Platform Challenges. *ACM Trans. Asian-Low-Resour. Lang. Inf. Process.* **2025**, *24*, 1–29. [CrossRef]
20. Patil, R.; Boit, S.; Gudivada, V.; Nandigam, J. A Survey of Text Representation and Embedding Techniques in NLP. *IEEE Access* **2023**, *11*, 36120–36146. [CrossRef]
21. Elmokhtar, G.; Mohamed, W.M.; El-Shafeiy, E.; Younis, E.M.G. Enhanced sentiment classification of phone brands on Twitter with a modified walrus optimizer and novel ensemble method. *Neural Comput. Appl.* **2025**, *37*, 14299–14325. [CrossRef]
22. Saifullah, S.; Dreżewski, R.; Dwiyanto, F.A.; Aribowo, A.S.; Fauziah, Y.; Cahyana, N.H. Automated Text Annotation Using a Semi-Supervised Approach with Meta Vectorizer and Machine Learning Algorithms for Hate Speech Detection. *Appl. Sci.* **2024**, *14*, 1078. [CrossRef]
23. Alsafari, S.; Sadaoui, S. Semi-Supervised Self-Training of Hate and Offensive Speech from Social Media. *Appl. Artif. Intell.* **2021**, *35*, 1621–1645. [CrossRef]
24. Neethu, M.S.; Vinod Chandra, S.S. A Review of Unlabeled and Imbalanced Data Challenges in Machine Learning: Strategies and Solutions. *WIREs Data Min. Knowl. Discov.* **2025**, *15*, e70043. [CrossRef]
25. Li, Z.; Ko, B.; Choi, H.J. Naive semi-supervised deep learning using pseudo-label. *Peer-Peer Netw. Appl.* **2019**, *12*, 1358–1368. [CrossRef]
26. Gui, Q.; Zhou, H.; Guo, N.; Niu, B. A survey of class-imbalanced semi-supervised learning. *Mach. Learn.* **2024**, *113*, 5057–5086. [CrossRef]
27. Cahyana, N.H.; Saifullah, S.; Fauziah, Y.; Aribowo, A.S.; Drezewski, R. Semi-supervised Text Annotation for Hate Speech Detection using K-Nearest Neighbors and Term Frequency-Inverse Document Frequency. *Int. J. Adv. Comput. Sci. Appl.* **2022**, *13*, 147–151. [CrossRef]
28. Mullah, N.S.; Zainon, W.M.N.W. Advances in Machine Learning Algorithms for Hate Speech Detection in Social Media: A Review. *IEEE Access* **2021**, *9*, 88364–88376. [CrossRef]
29. Naseeb, A.; Zain, M.; Hussain, N.; Qasim, A.; Ahmad, F.; Sidorov, G.; Gelbukh, A. Machine Learning- and Deep Learning-Based Multi-Model System for Hate Speech Detection on Facebook. *Algorithms* **2025**, *18*, 331. [CrossRef]
30. Saifullah, S.; Fauziyah, Y.; Aribowo, A.S. Comparison of machine learning for sentiment analysis in detecting anxiety based on social media data. *J. Inform.* **2021**, *15*, 45. [CrossRef]
31. Marshan, A.; Nizar, F.N.M.; Ioannou, A.; Spanaki, K. Comparing Machine Learning and Deep Learning Techniques for Text Analytics: Detecting the Severity of Hate Comments Online. *Inf. Syst. Front.* **2025**, *27*, 487–505. [CrossRef]
32. Allam, H.; Makubvure, L.; Gyamfi, B.; Graham, K.N.; Akinwolere, K. Text Classification: How Machine Learning Is Revolutionizing Text Categorization. *Information* **2025**, *16*, 130. [CrossRef]
33. Mubeen, M.; Muskan, A.; Akram, A.; Rashid, J.; Alshalali, T.A.N.; Sarwar, N. Cyberbullying-Related Automated Hate Speech Detection on Social Media Platforms Using Stack Ensemble Classification Method. *Int. J. Comput. Intell. Syst.* **2025**, *18*, 174. [CrossRef]
34. Goswami, P.; Daniel, A. Enhancing Hate Speech Detection with a Distil BERT and BiLSTM Hybrid Mode. *SN Comput. Sci.* **2026**, *7*, 353. [CrossRef]
35. Hashmi, E.; Ahmad, H.; Mazhar, M.T.; Yayilgan, S.Y.; Afzal, M.; Shaikh, S. A Robust and Linguistically-Aware Hate Speech Detection System for Roman Urdu. *ACM Trans. Asian-Low-Resour. Lang. Inf. Process.* **2025**, *24*, 1–22. [CrossRef]
36. Gasmi, K.; Ben Ltaifa, I.; Eltoum Abdalrahman, A.; Hamid, O.; Othman Altaieb, M.; Ali, S.; Ben Ammar, L.; Mrabet, M. Hybrid Feature and Optimized Deep Learning Model Fusion for Detecting Hateful Arabic Content. *IEEE Access* **2025**, *13*, 131411–131431. [CrossRef]

37. Bolatbek, M.; Sagynay, M.; Mussiraliyeva, S.; Yeltay, Z. Detection of offensive content in the Kazakh language using machine learning and deep learning approaches. *PeerJ Comput. Sci.* **2025**, *11*, e3027. [CrossRef] [PubMed]
38. Geetanjali.; Kumar, M. Exploring hate speech detection: Challenges, resources, current research and future directions. *Multimed. Tools Appl.* **2025**, *84*, 38423–38459. [CrossRef]
39. Gandhi, A.; Ahir, P.; Adhvaryu, K.; Shah, P.; Lohiya, R.; Cambria, E.; Poria, S.; Hussain, A. Hate speech detection: A comprehensive review of recent works. *Expert Syst.* **2024**, *41*, e13562. [CrossRef]
40. Alghamdi, J.; Lin, Y.; Luo, S. Machine learning and deep learning approaches for fake news detection and related topics in multilingual contexts: A systematic literature review. *Multimed. Tools Appl.* **2026**, *85*, 353. [CrossRef]
41. Ngueajio, M.K.; Aryal, S.; Atemkeng, M.; Washington, G.; Rawat, D. Decoding Fake News and Hate Speech: A Survey of Explainable AI Techniques. *ACM Comput. Surv.* **2025**, *57*, 1–37. [CrossRef] [PubMed]
42. Pamungkas, E.W.; Chiril, P. Ngalawan Ujaran Sengit: Hate speech detection in indonesian code-mixed social media data. *Lang. Resour. Eval.* **2025**, *59*, 2387–2414. [CrossRef]
43. Ibrohim, M.O.; Budi, I. Hate speech and abusive language detection in Indonesian social media: Progress and challenges. *Heliyon* **2023**, *9*, e18647. [CrossRef] [PubMed]
44. Poletto, F.; Basile, V.; Sanguinetti, M.; Bosco, C.; Patti, V. Resources and benchmark corpora for hate speech detection: A systematic review. *Lang. Resour. Eval.* **2021**, *55*, 477–523. [CrossRef]
45. Narula, R.; Chaudhary, P. A comprehensive review on detection of hate speech for multi-lingual data. *Soc. Netw. Anal. Min.* **2025**, *14*, 244. [CrossRef]
46. Pawar, S.; Park, J.; Jin, J.; Arora, A.; Myung, J.; Yadav, S.; Haznitrama, F.G.; Song, I.; Oh, A.; Augenstein, I. Survey of Cultural Awareness in Language Models: Text and Beyond. *Comput. Linguist.* **2025**, *51*, 907–1004. [CrossRef]
47. Nkemelu, D.; Shah, H.; Best, M.; Essa, I. Tackling Hate Speech in Low-resource Languages with Context Experts. In Proceedings of the International Conference on Information & Communication Technologies and Development 2022, New York, NY, USA, 27–29 June 2022; pp. 1–11. [CrossRef]
48. Qureshi, M.D.M.; Qureshi, M.A.; Rashwan, W. Explainable AI for Hate Speech Moderation: A Stakeholder-Centered and Sociotechnical Review. *WIREs Data Min. Knowl. Discov.* **2026**, *16*, e70076. [CrossRef]
49. Afkari-Fahandari, A.; Shabaninia, E.; Asadi-Zeydabadi, F.; Nezamabadi-Pour, H. A Comprehensive Survey of Transformers in Text Recognition: Techniques, Challenges, and Future Directions. *ACM Comput. Surv.* **2026**, *58*, 1–42. [CrossRef]
50. Yu, J.S.; Yao, Y. Language Modeling and Large Language Models. In *Intelligent Language Services*; Springer Nature: Singapore, 2026; pp. 33–62. [CrossRef]
51. Mostafavi, S.; Yahyavi, Y.; Ravanmehr, R. Systematic literature review on sentiment analysis using transformers. *Int. J. Data Sci. Anal.* **2026**, *22*, 49. [CrossRef]
52. Li, Q.; Peng, H.; Li, J.; Xia, C.; Yang, R.; Sun, L.; Yu, P.S.; He, L. A Survey on Text Classification: From Traditional to Deep Learning. *ACM Trans. Intell. Syst. Technol.* **2022**, *13*, 1–41. [CrossRef]
53. Cunha, W.; Viegas, F.; França, C.; Rosa, T.; Rocha, L.; Gonçalves, M.A. A Comparative Survey of Instance Selection Methods applied to Non-Neural and Transformer-Based Text Classification. *ACM Comput. Surv.* **2023**, *55*, 1–52. [CrossRef]
54. Hidayatullah, A.F.; Apong, R.A.; Lai, D.T.C.; Qazi, A. Pre-trained language model for code-mixed text in Indonesian, Javanese, and English using transformer. *Soc. Netw. Anal. Min.* **2025**, *15*, 30. [CrossRef]
55. Jayadianti, H.; Kaswidjanti, W.; Utomo, A.T.; Saifullah, S.; Dwiyanto, F.A.; Drezewski, R. Sentiment analysis of Indonesian reviews using fine-tuning IndoBERT and R-CNN. *ILKOM J. Ilm.* **2022**, *14*, 348–354. [CrossRef]
56. Ridha, M.; Nurjanah, D.; Rakha, M. Multilabel Classification Abusive Language and Hate Speech on Indonesian Twitter Using Transformer Model: IndoBERTweet & IndoRoBERTa. In *Proceedings of the 2024 International Conference on Intelligent Cybernetics Technology & Applications (ICICyTA)*; IEEE: Piscataway, NJ, USA, 2024; pp. 48–54. [CrossRef]
57. Pramana, R.; Jonathan, M.; Yani, H.S.; Sutoyo, R. A Comparison of BiLSTM, BERT, and Ensemble Method for Emotion Recognition on Indonesian Product Reviews. *Procedia Comput. Sci.* **2024**, *245*, 399–408. [CrossRef]
58. Awal, M.R.; Lee, R.K.W.; Tanwar, E.; Garg, T.; Chakraborty, T. Model-Agnostic Meta-Learning for Multilingual Hate Speech Detection. *IEEE Trans. Comput. Soc. Syst.* **2024**, *11*, 1086–1095. [CrossRef]
59. Xue, Q.; Dou, Y.; Shi, Z.R.; Li, X.L.; Gao, W. MMBERT: Scaled Mixture-of-Experts Multimodal BERT for Robust Chinese Hate Speech Detection Under Cloaking Perturbations. *Proc. AAAI Conf. Artif. Intell.* **2026**, *40*, 34196–34204. [CrossRef]
60. Depuru, S.; Mahathi, T.; Prasad, V.V.; Devi Daswanth, S.; Rohini, T. Cross-Lingual Hate Speech Detection using XLM-R and Transformer based Classification. In *Proceedings of the 2025 International Conference on Sustainable Communication Networks and Application (ICSCN)*; IEEE: Piscataway, NJ, USA, 2025; pp. 1481–1486. [CrossRef]
61. Nurrahmi, H.; Wibowo, A.T.; Hasmawati. Evaluating Classification Models Over IndoBERT and IndoBERTweet Embeddings for Indonesian Cyberbullying Detection. In *Proceedings of the 2025 5th International Conference on Intelligent Cybernetics Technology & Applications (ICICyTA)*; IEEE: Piscataway, NJ, USA, 2025; pp. 268–273. [CrossRef]

62. Tabinda Kokab, S.; Asghar, S.; Naz, S. Transformer-based deep learning models for the sentiment analysis of social media data. *Array* **2022**, *14*, 100157. [CrossRef]
63. Ramos, G.; Batista, F.; Ribeiro, R.; Fialho, P.; Moro, S.; Fonseca, A.; Guerra, R.; Carvalho, P.; Marques, C.; Silva, C. A comprehensive review on automatic hate speech detection in the age of the transformer. *Soc. Netw. Anal. Min.* **2024**, *14*, 204. [CrossRef]
64. Pan, R.; García-Díaz, J.A.; Valencia-García, R. Comparing Fine-Tuning, Zero and Few-Shot Strategies with Large Language Models in Hate Speech Detection in English. *Comput. Model. Eng. Sci.* **2024**, *140*, 2849–2868. [CrossRef]
65. Diaz-Garcia, J.A.; Carvalho, J.P. A Literature Review of Textual Cyber Abuse Detection Using Cutting-Edge Natural Language Processing Techniques: Language Models and Large Language Models. *WIREs Data Min. Knowl. Discov.* **2025**, *15*, e70029. [CrossRef]
66. Findawati, Y.; Budi Raharjo, A.; Adni Navastara, D.; Yonathan, V.; Yatestha, A.A.; Purwitasari, D. Multi-label Aspect Dangerous Speech Classification Using Keyword-Driven Ensemble Classifier on Imbalanced Data. *JOIV Int. J. Inform. Vis.* **2025**, *9*, 1592. [CrossRef]
67. Prabhu, R.; Seethalakshmi, V. A comprehensive framework for multi-modal hate speech detection in social media using deep learning. *Sci. Rep.* **2025**, *15*, 13020. [CrossRef] [PubMed]
68. Guleria, P.; Frnda, J.; Srinivasu, P.N. NLP based text classification using TF-IDF enabled fine-tuned long short-term memory: An empirical analysis. *Array* **2025**, *27*, 100467. [CrossRef]
69. Yadav, A.; Singh, V. HateFusion: Harnessing Attention-Based Techniques for Enhanced Filtering and Detection of Implicit Hate Speech. *IEEE Trans. Comput. Soc. Syst.* **2025**, *12*, 1700–1715. [CrossRef]
70. Roy, P.K. MMFFHS: Multi-Modal Feature Fusion for Hate Speech Detection on Social Media. *IEEE Trans. Big Data* **2025**, *11*, 1247–1258. [CrossRef]
71. Dwivedy, V.; Roy, P.K. Deep feature fusion for hate speech detection: A transfer learning approach. *Multimed. Tools Appl.* **2023**, *82*, 36279–36301. [CrossRef]
72. Ariyanto, A.D.P.; Purwitasari, D.; Fatichah, C. A Systematic Review on Semantic Role Labeling for Information Extraction in Low-Resource Data. *IEEE Access* **2024**, *12*, 57917–57946. [CrossRef]
73. Alkılınç, A.; Yıldırım Okay, F.; Kök, I.; Özdemir, S. A comprehensive survey on federated semi-supervised learning: applications, challenges, and future directions. *Clust. Comput.* **2026**, *29*, 221. [CrossRef]
74. Malik, U.; Bernard, S.; Pauchet, A.; Chatelain, C.; Picot-Clémente, R.; Cortinovis, J. Pseudo-Labeling With Large Language Models for Multi-Label Emotion Classification of French Tweets. *IEEE Access* **2024**, *12*, 15902–15916. [CrossRef]
75. Dirting, B.D.; Chukwudebe, G.A.; Nwokorie, E.C.; Ayogu, I.I. Multi-Label Classification of Hate Speech Severity on Social Media using BERT Model. In *Proceedings of the 2022 IEEE Nigeria 4th International Conference on Disruptive Technologies for Sustainable Development (NIGERCON)*; IEEE: Piscataway, NJ, USA, 2022; pp. 1–5. [CrossRef]
76. Gui, J.; Chen, T.; Zhang, J.; Cao, Q.; Sun, Z.; Luo, H.; Tao, D. A Survey on Self-Supervised Learning: Algorithms, Applications, and Future Trends. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *46*, 9052–9071. [CrossRef] [PubMed]
77. Abdulrazzaq, M.M.; Ramaha, N.T.A.; Hameed, A.A.; Salman, M.; Yon, D.K.; Fitriyani, N.L.; Syafrudin, M.; Lee, S.W. Consequential Advancements of Self-Supervised Learning (SSL) in Deep Learning Contexts. *Mathematics* **2024**, *12*, 758. [CrossRef]
78. Pannerselvam, K.; Rajiakodi, S. Systematic literature review on hate speech detection in Indian low-resource languages. *J. Comput. Soc. Sci.* **2026**, *9*, 5. [CrossRef]
79. Mane, S.S.; Kundu, S.; Sharma, R. A Survey on Online Aggression: Content Detection and Behavioral Analysis on Social Media. *ACM Comput. Surv.* **2025**, *57*, 1–36. [CrossRef]
80. Pakray, P.; Gelbukh, A.; Bandyopadhyay, S. Natural language processing applications for low-resource languages. *Nat. Lang. Process.* **2025**, *31*, 183–197. [CrossRef]
81. Ronny Mabokela, K.; Primus, M.; Celik, T. Advancing sentiment analysis for low-resourced african languages using pre-trained language models. *PLoS ONE* **2025**, *20*, e0325102. [CrossRef] [PubMed]
82. Khandokar, I.A.; Deshpande, P. SSABE-TSCM: Drift-aware and interpretable financial sentiment analysis for low-resource Bangla via adaptive semi-supervised and temporal contrastive modeling. *Front. Artif. Intell.* **2026**, *9*, 1724407. [CrossRef] [PubMed]
83. Taha, K. Big Data Analytics in IoT, social media, NLP, and information security: Trends, challenges, and applications. *J. Big Data* **2025**, *12*, 150. [CrossRef]
84. Sarzaeim, P.; Azim, A.; Bauer, G.; Makrehchi, M. A Survey of Domain-Specific Fine-Tuned Large Language Models. *IEEE Access* **2026**, *14*, 48407–48433. [CrossRef]
85. Da Ros, F.; Soprano, M.; Di Gaspero, L.; Roitero, K. Large Language Models for Combinatorial Optimization: A Systematic Review. *ACM Comput. Surv.* **2026**, *58*, 1–53. [CrossRef]
86. Usman, M.; Ahmad, M.; Sidorov, G.; Gelbukh, I.; Tellez, R.Q. A Large Language Model-Based Approach for Multilingual Hate Speech Detection on Social Media. *Computers* **2025**, *14*, 279. [CrossRef]
87. Mullah, N.S.; Wan Zainon, W.M.N.; Ridzuan, F.; Abiodun, E.O. Machine learning intervention on cyber-hate in code-switch texts: A systematic review with open challenges and solutions. *PeerJ Comput. Sci.* **2026**, *12*, e3537. [CrossRef]

88. Aribowo, A.S.; Khomsah, S.; Saifullah, S. Semi-Supervised Sentiment Classification Using Self-Learning and Enhanced Co-Training. *J. Infotel* **2025**, *17*, 472–489. [CrossRef]
89. Mujahid, M.; Kına, E.; Rustam, F.; Villar, M.G.; Alvarado, E.S.; De La Torre Diez, I.; Ashraf, I. Data oversampling and imbalanced datasets: An investigation of performance for machine learning and feature engineering. *J. Big Data* **2024**, *11*, 87. [CrossRef]
90. Kumar S., Y.P.; Mishra, S. Enhancing network traffic classification with autoencoders and TF–IDF-based dimensionality reduction: A data-centric comparative study. *Comput. Netw.* **2026**, *282*, 112306. [CrossRef]
91. Ahmed, T.; Ivan, S.; Kabir, M.; Mahmud, H.; Hasan, K. Performance analysis of transformer-based architectures and their ensembles to detect trait-based cyberbullying. *Soc. Netw. Anal. Min.* **2022**, *12*, 99. [CrossRef]
92. OpenAI. GPT-4o Mini: Advancing Cost-Efficient Intelligence. 2024. Available online: https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/ (accessed on 12 May 2026).
93. Grattafiori, A.; Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Vaughan, A.; et al. The Llama 3 Herd of Models. *arXiv* **2024**, arXiv:2407.21783. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.