KETIK : Jurnal Informatika
ISSN: 3026-183X (Media Online)
Vol. 03, No. 06, Juli 2026 Hal. 275-283
Publisher: Faatuatua Media Karya
jurnal.faatuatua.com/index.php/KETIK
DOI: doi.org/10.70404/ketik.v3i06.697

Perbandingan CNN-BiLSTM dan CNN-BiGRU untuk Klasifikasi Berita Indonesia Menggunakan Fast Text Embedding dan Confidence Based Pseudo Labeling

Cindy Rahmayanti[1], Elisa Wulan Sari[2*], Adha Mashur Sajiah[3]
[1]*Teknik Informatika, Universitas Halu Oleo, Indonesia, e-mail: **cindyrahmayanti43@gmail.com***
[2*]*Teknik Informatika, Universitas Halu Oleo, Indonesia, e-mail: **elisawsari06@gmail.com***
[3]*Teknik Informatika, Universitas Halu Oleo, Indonesia, e-mail: **adha.m.sajiah@uho.ac.id***
**coressponding author)*

Info Artikel

Diajukan: *23-06-2026*
Direvisi: *03-07-2026*
Diterima: *29-07-2026*
Diterbitkan: *Juli 2026*

Kata Kunci:
Klasifikasi Teks;
CNN; LSTM;
GRU;
FastText;
Pseudo Labeling.

Keywords:
Text Classification;
CNN;
LSTM;
GRU;
FastText;
Pseudo-Labeling.

Abstrak

Klasifikasi teks berita Indonesia merupakan tugas penting dalam pemrosesan bahasa alami yang mendukung pengelolaan informasi di era digital. Penelitian ini mengusulkan perbandingan dua arsitektur deep learning, yaitu CNN-BiLSTM dan CNN-BiGRU, yang diintegrasikan dengan word embedding FastText bahasa Indonesia dan strategi semi-supervised learning berbasis Confidence-Based Pseudo-Labeling. Dataset yang digunakan adalah Indonesian News Dataset sebanyak 91.017 artikel dari 9 kategori, yang kemudian diseimbangkan menggunakan undersampling menjadi 21.924 sampel. Sebanyak 20% data digunakan sebagai data berlabel dan 80% sisanya diperlakukan sebagai data tak berlabel yang dimanfaatkan melalui tiga iterasi pseudo-labeling dengan ambang kepercayaan adaptif (0,85; 0,88; 0,90). Hasil evaluasi menunjukkan menunjukkan bahwa penerapan pseudo-labeling secara konsisten meningkatkan performa kedua model. CNN-BiLSTM mencapai akurasi 71,72% dan F1-score 71,60% setelah pseudo-labeling, sedangkan CNN-BiGRU mencapai akurasi 69,42% dan F1-score 69,19%. Temuan ini membuktikan bahwa pendekatan semi-supervised learning efektif digunakan pada kondisi data berlabel yang terbatas

Abstract

Indonesian news text classification is an important task in natural language processing that supports information management in the digital era. This study proposes a comparison of two deep learning architectures, CNN-BiLSTM and CNN-BiGRU, integrated with Indonesian FastText word embeddings and a semi-supervised learning strategy based on Confidence-Based Pseudo-Labeling. The dataset used is the Indonesian News Dataset consisting of 91,017 articles from 9 categories, balanced via undersampling to 21,924 samples. Twenty percent of the data were used as labeled data, while the remaining 80% served as unlabeled data exploited through three pseudo-labeling iterations with adaptive confidence thresholds (0.85, 0.88, 0.90). Evaluation results demonstrate that pseudo-labeling consistently improves the performance of both models. CNN-BiLSTM achieves 71.72% accuracy and 71.60% F1-score after pseudo-labeling, while CNN-BiGRU reaches 69.42% accuracy and 69.19% F1-score. These findings confirm that semi-supervised learning is effective when labeled data is scarce.

1. PENDAHULUAN

Pertumbuhan platform berita digital di Indonesia menghasilkan volume konten yang terus meningkat setiap harinya. Ribuan artikel dipublikasikan dari berbagai kanal media daring, mencakup topik yang beragam mulai dari politik, ekonomi, olahraga, hingga teknologi. Kondisi ini mendorong kebutuhan nyata akan sistem yang mampu mengkategorikan berita secara otomatis dan akurat, sehingga membantu pengguna dalam menyaring dan menemukan informasi yang relevan secara efisien [1].

Klasifikasi teks merupakan salah satu tugas fundamental dalam *Natural Language Processing* (NLP) yang bertujuan mengelompokkan dokumen teks ke dalam kategori yang telah ditentukan. Dalam

konteks berita berbahasa Indonesia, tantangan utama yang dihadapi meliputi kekayaan morfologi bahasa Indonesia, variasi diksi jurnalistik antar media, serta ketidakseimbangan distribusi kategori berita. Pendekatan *deep learning* telah terbukti unggul dibandingkan metode *machine learning* konvensional untuk tugas pemrosesan teks, terutama karena kemampuannya dalam mempelajari representasi fitur secara hierarkis [2].

Arsitektur *Convolutional Neural Network* (CNN) dan *Recurrent Neural Network* (RNN) berbasis *Long Short-Term Memory* (LSTM) serta *Gated Recurrent Unit* (GRU) telah banyak diaplikasikan pada klasifikasi teks. Penelitian Guridno et al. [3] menunjukkan bahwa model hibrid CNN-LSTM mampu meningkatkan akurasi klasifikasi topik berita Indonesia secara signifikan dibandingkan metode tunggal. Sementara itu, Hanif et al. [4] membandingkan kinerja LSTM, Bi-LSTM, dan GRU pada klasifikasi judul berita *clickbait* dan menemukan bahwa arsitektur bidireksional memberikan peningkatan performa yang konsisten. Arsitektur bidireksional memungkinkan model membaca urutan teks dari dua arah sehingga dapat menangkap konteks yang lebih kaya [5].

Representasi teks merupakan komponen kritis dalam klasifikasi teks berbasis *deep learning*. Nurdin et al. [6] membandingkan Word2Vec, GloVe, dan FastText pada klasifikasi teks dan menemukan bahwa FastText unggul dengan *F-measure* 0,979 pada dataset 20 Newsgroup. Keunggulan FastText terletak pada kemampuannya merepresentasikan kata menggunakan sub-kata (*subword*) sehingga mampu menangani kata yang belum pernah ditemukan sebelumnya maupun kata berimbuhan yang umum dalam bahasa Indonesia [7].

Keterbatasan data berlabel merupakan kendala nyata dalam pengembangan model klasifikasi teks. Proses anotasi manual membutuhkan waktu dan biaya yang besar, sehingga *semi-supervised learning* (SSL) menjadi solusi yang menarik karena memanfaatkan data tak berlabel yang tersedia melimpah [8]. Teknik *pseudo-labeling* merupakan salah satu pendekatan SSL yang paling efisien, di mana model yang sudah dilatih pada data berlabel digunakan untuk memberikan label sementara pada data tak berlabel berdasarkan ambang kepercayaan tertentu [9].

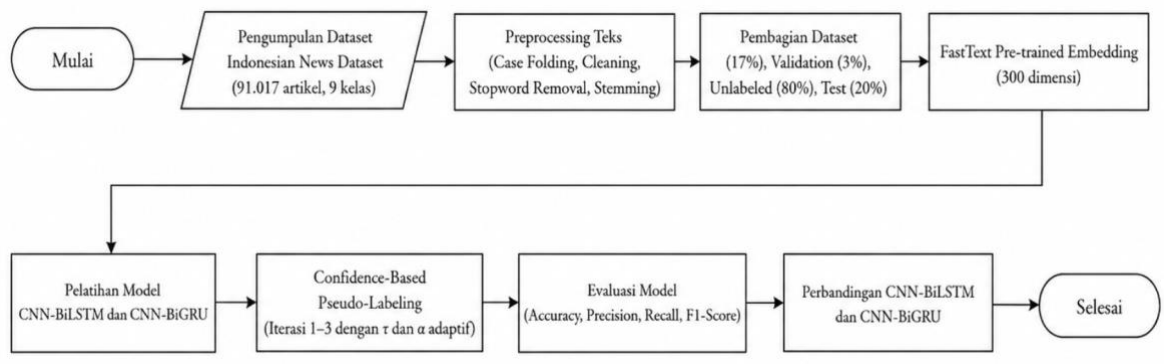
Penelitian ini mengusulkan pendekatan yang menggabungkan kekuatan beberapa komponen: (1) arsitektur hybrid CNN-BiLSTM dan CNN-BiGRU yang memanfaatkan kemampuan CNN dalam ekstraksi fitur lokal dan kemampuan jaringan berulang bidireksional dalam menangkap konteks sekuensial, (2) pre-trained FastText Bahasa Indonesia sebagai representasi kata yang kaya secara semantik dan mampu menangani kata di luar kosakata (OOV) melalui subword information , serta (3) mekanisme Confidence-Based Pseudo-Labeling dengan threshold adaptif yang hanya memanfaatkan prediksi dengan tingkat keyakinan tinggi, sehingga meminimalkan propagasi label yang keliru.

Beberapa penelitian terdahulu telah mengeksplorasi klasifikasi berita berbahasa Indonesia menggunakan deep learning [10]. Namun, sebagian besar penelitian tersebut masih mengandalkan pendekatan supervised learning penuh yang membutuhkan data berlabel dalam jumlah besar, dan belum memanfaatkan FastText sebagai embedding sekaligus mengombinasikannya dengan mekanisme semi-supervised learning. Penelitian ini mengisi celah tersebut dengan secara sistematis membandingkan dua varian arsitektur, mengevaluasi kontribusi pseudo-labeling secara kuantitatif, serta menggunakan dataset yang telah diseimbangkan untuk memastikan evaluasi yang adil antar kelas.

Tujuan penelitian ini adalah (1) membangun dan melatih model CNN-BiLSTM dan CNN-BiGRU menggunakan FastText embedding, (2) menerapkan mekanisme Confidence-Based Pseudo-Labeling untuk memanfaatkan data tak berlabel, dan (3) menganalisis perbandingan performa keempat skenario secara komprehensif menggunakan metrik akurasi, presisi, recall, dan F1-score.

2. METODE PENELITIAN

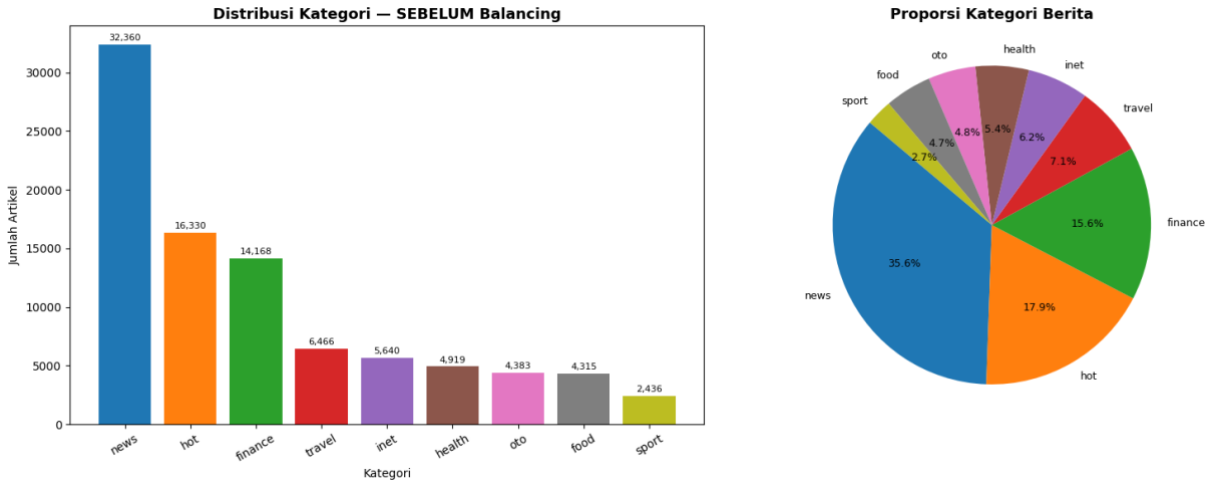
Penelitian ini mengikuti alur metodologi yang terdiri dari enam tahap utama: pengumpulan dan eksplorasi data, praproses teks, pembagian data, pembangunan embedding FastText, pelatihan model, dan evaluasi. Kerangka kerja implementasi menggunakan TensorFlow/Keras pada lingkungan Google Colaboratory.



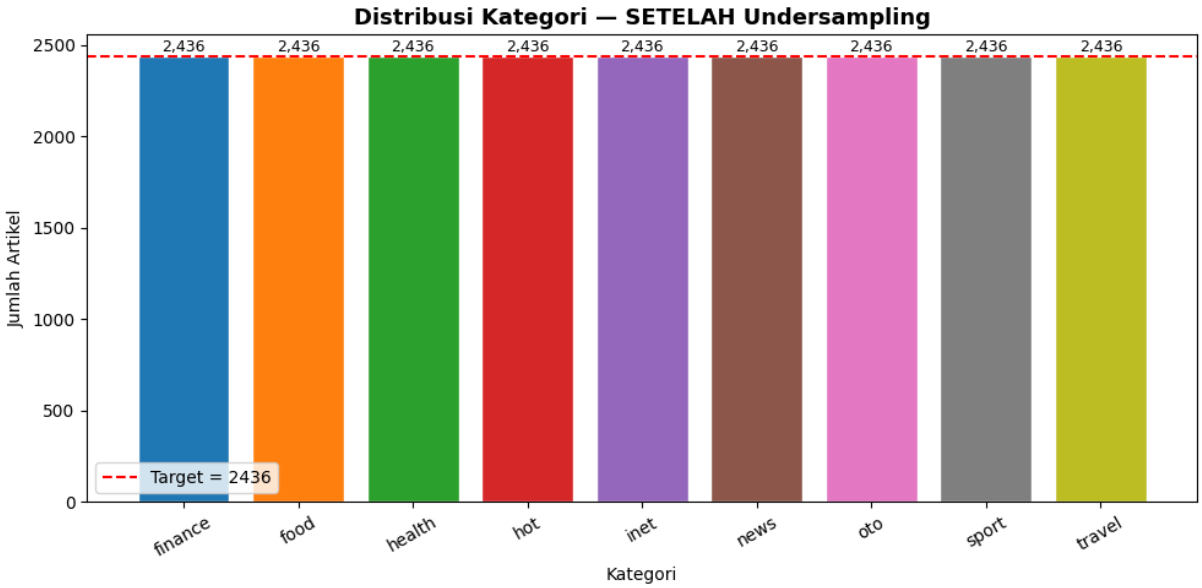
Gambar 1. Alur metodologi penelitian

2.1 *Dataset dan Pra-pemrosesan*

Dataset yang digunakan dalam penelitian ini adalah Indonesian News Dataset yang di peroleh dari platform Kaggle. Dataset ini terdiri dari 91.017 artikel berita berbahasa indonesia yang bersumber dari portal berita detik.com Setiap data telah dilengkapi dengan label kategori yang mencakup sembilan kelas, yaitu news, hot, finance, travel, inet, health, oto, food, dan sport. Mengingat distribusi kelas yang tidak seimbang pada dataset asli, dilakukan proses undersampling dengan mengambil 2.436 sampel dari setiap kelas, sehingga total dataset setelah penyeimbangan berjumlah 21.924 sampel. Keputusan ini didasarkan pada jumlah sampel kelas minoritas (sport). Visualisasi distribusi data sebelum dan sesudah balancing ditunjukkan pada Gambar 2 dan Gambar 3.



Gambar 2. Distribusi kategori berita sebelum undersampling (91.017 artikel)



Gambar 3. Distribusi kategori berita setelah undersampling (21.924 artikel, 2.436 per kelas)

Tahapan pra-pemrosesan teks menggunakan library PySastrawi untuk bahasa Indonesia dan mencakup: (1) normalisasi teks (*lowercase*, penghapusan URL, angka, dan tanda baca), (2) penghapusan *stopword* bahasa Indonesia (356 kata), dan (3) stemming menggunakan algoritma Sastrawi. Panjang maksimum sekuens ditetapkan 50 token, yang mencakup 96,8% judul berita dalam dataset. Ukuran kosakata dibatasi 30.000 kata.

2.2 Pembagian Data

Data dibagi secara bertingkat (*stratified*) untuk menjaga proporsi kelas. Pembagian data ditunjukkan pada Tabel 1.

Tabel 1. Pembagian dataset penelitian

Subset Data	Jumlah Sampel	Persentase	Keterangan
Data Berlabel (Labeled)	4.384	20%	Train + Validasi
Data Latih (Train)	3.726	17%	Supervised training
Data Validasi	658	3%	Model selection
Data Tak Berlabel (Unlabeled)	13.155	60%	Semi-supervised
Data Uji (Test)	4.385	20%	Evaluasi akhir
Total Dataset Balanced	21.924	100%	9 kelas, 2.436/kelas

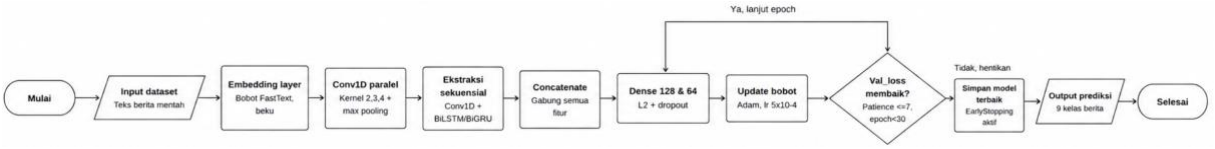
2.3 *FastText Pre-trained Embedding*

Word embedding diinisialisasi menggunakan model FastText bahasa Indonesia yang telah dilatih sebelumnya (*pre-trained*) dari Facebook AI Research (cc.id.300.vec), dengan dimensi vektor 300. Model ini dilatih pada korpus Wikipedia dan CommonCrawl dalam bahasa Indonesia, mencakup jutaan kata termasuk kata berimbuhan [11]. Matriks embedding diinisialisasi untuk kosakata yang cocok antara tokenizer dan model FastText, sementara kata yang tidak ditemukan diinisialisasi secara acak dari distribusi normal kecil. Coverage FastText pada kosakata model mencapai lebih dari 85% dari total kosakata aktif (VOCAB_SIZE = 30.000).

2.4 *Arsitektur Model (gambar arsitekturnya)*

Kedua model berbagi desain arsitektur yang serupa, perbedaannya hanya pada lapisan *recurrent*. Setiap model terdiri dari: (1) lapisan *Embedding* yang diinisialisasi dengan bobot FastText (tidak dilatih ulang), (2) tiga lapisan *Conv1D* paralel dengan ukuran kernel 2, 3, dan 4 (masing-masing 128 filter) dilanjutkan *GlobalMaxPooling1D*, (3) dua lapisan *Conv1D* bertumpuk untuk mengekstraksi fitur sekuensial, (4) lapisan *Bidirectional LSTM/GRU* dengan 128 unit, (5) penggabungan (*concatenate*) keluaran semua cabang, dan (6) dua lapisan Dense (128 dan 64 unit) dengan regularisasi L2 dan Dropout sebelum lapisan Softmax output (9 kelas). *Hyperparameter* pelatihan mencakup optimizer

Adam (learning rate 5×10^{-4}), loss categorical cross-entropy, batch size 64, dan maksimum 30 epoch dengan EarlyStopping (patience=7).



Gambar 4. Arsitektur Model

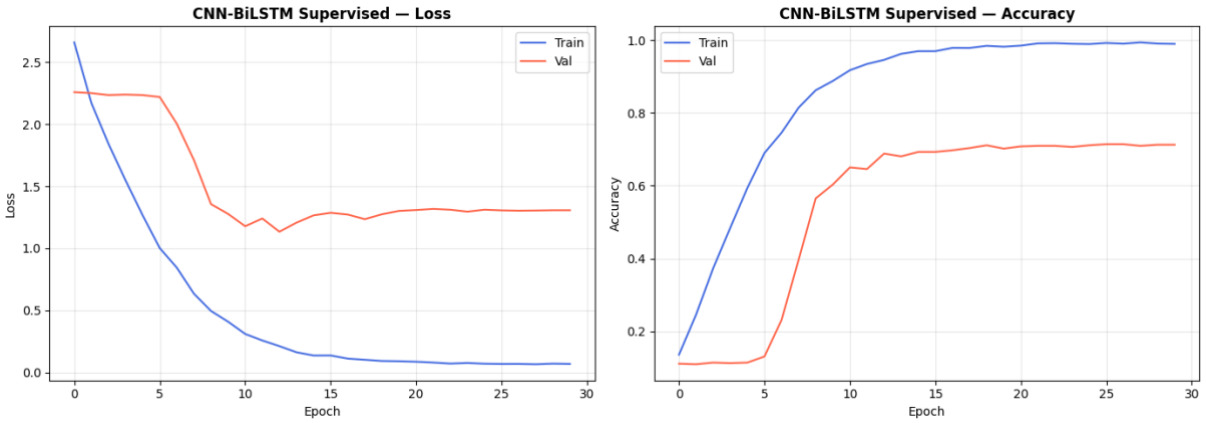
2.5 *Confidence-Based Pseudo-Labeling*

Strategi semi-supervised yang digunakan adalah *Confidence-Based Pseudo-Labeling* dengan tiga iterasi. Pada setiap iterasi, model memprediksi probabilitas kelas untuk data tak berlabel, kemudian hanya sampel dengan kepercayaan maksimum (max softmax probability) melebihi ambang τ yang dipilih sebagai pseudo-label. Untuk mengurangi efek label yang berisik, pseudo-label diberi bobot α yang meningkat secara jadwal (α = 0,3 → 0,6 → 1,0) pada setiap iterasi, sehingga kontribusi pseudo-label bertambah seiring meningkatnya kualitas model [12]. Nilai ambang juga dinaikkan secara progresif: τ = 0,85 (iterasi 1), τ = 0,88 (iterasi 2), τ = 0,90 (iterasi 3). Peningkatan ambang secara bertahap ini bertujuan untuk memastikan bahwa hanya sampel dengan tingkat keyakinan yang benar-benar tinggi yang dimasukkan ke dalam proses pelatihan pada iterasi lanjut, mengingat model pada iterasi awal masih rentan menghasilkan prediksi yang kurang akurat. Dengan kombinasi strategi peningkatan bobot α dan ambang τ tersebut, risiko akumulasi kesalahan (error propagation) akibat pseudo-label yang keliru dapat ditekan, sehingga performa model secara keseluruhan menjadi lebih stabil pada iterasi berikutnya.

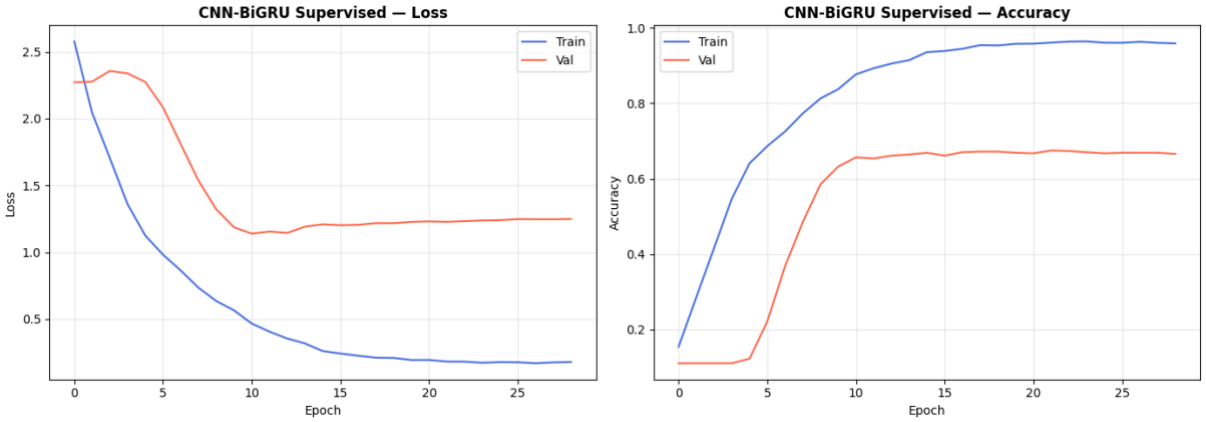
3. HASIL DAN ANALISIS

3.1 Evaluasi Model Supervised (Baseline)

Pada tahap baseline, kedua model dilatih hanya menggunakan 4.384 sampel berlabel (20% dataset). Selama pelatihan, EarlyStopping menghentikan pelatihan sebelum mencapai 30 epoch untuk mencegah *overfitting*. Kurva pelatihan CNN-BiLSTM dan CNN-BiGRU ditunjukkan pada Gambar 6 dan 7. Hasil evaluasi pada data uji disajikan dalam Tabel 2.



Gambar 4. Kurva loss dan akurasi pelatihan CNN-BiLSTM (supervised)



Gambar 5. Kurva loss dan akurasi pelatih pelatihan CNN-BiGRU (supervised)

Tabel 2. Perbandingan hasil evaluasi keempat skenario model pada data uji

Model / Skenario	Akurasi	Precision	Recall	F1
CNN-BiLSTM Supervised	0.6965	0.7154	0.6965	0.6994
CNN-BiLSTM + Pseudo-Labeling	0.7172	0.7284	0.7172	0.7160
CNN-BiGRU Supervised	0.6862	0.7059	0.6862	0.6881
CNN-BiGRU + Pseudo-Labeling	0.6942	0.7091	0.6942	0.6919

Tabel 2 menunjukkan bahwa model yang menggunakan Pseudo-Labeling memperoleh hasil yang lebih baik dibandingkan model supervised. Pada CNN-BiLSTM, akurasi meningkat dari 69,65% menjadi 71,72%, sedangkan F1-score meningkat dari 69,94% menjadi 71,60%. Pada CNN-BiGRU, akurasi meningkat dari 68,62% menjadi 69,42% dan F1-score dari 68,81% menjadi 69,19%. Hasil ini menunjukkan bahwa penerapan Confidence-Based Pseudo-Labeling mampu meningkatkan performa model dengan memanfaatkan data tak berlabel sebagai tambahan data pelatihan.

3.2 Statistik Pseudo-Labeling

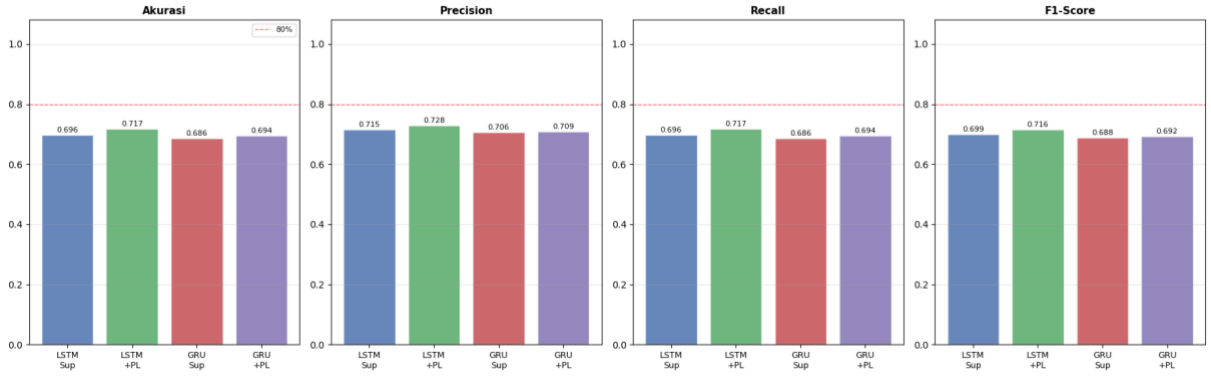
Selama proses pseudo-labeling, jumlah data tak berlabel yang berhasil diberikan label sementara pada setiap iterasi dicatat untuk kedua model. Tabel 3 merangkum statistik pseudo-labeling, termasuk jumlah sampel yang diterima (confidence ≥ τ) dan akurasi validasi setelah setiap iterasi.

Tabel 3. Statistik confidence-based pseudo-labeling per iterasi

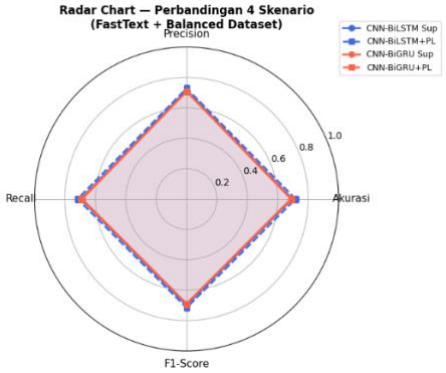
Iterasi	T	α	Pseudo-Label (LSTM)	Pseudo-Label (GRU)	Val Acc (GRU)
1	0,85	0,3	6846	4879	0,702
2	0,88	0,6	2830	2281	0,6991
3	0,90	1,0	327	358	0,6976

3.3 Analisis Komparatif

Gambar 8 menampilkan perbandingan keempat metrik evaluasi (akurasi, precision, recall, dan F1-Score) untuk keempat skenario dalam bentuk *bar chart*, semenara Gambar 9 menampilkan *radar chart* untuk visualisasi perbandingan komprehensif.

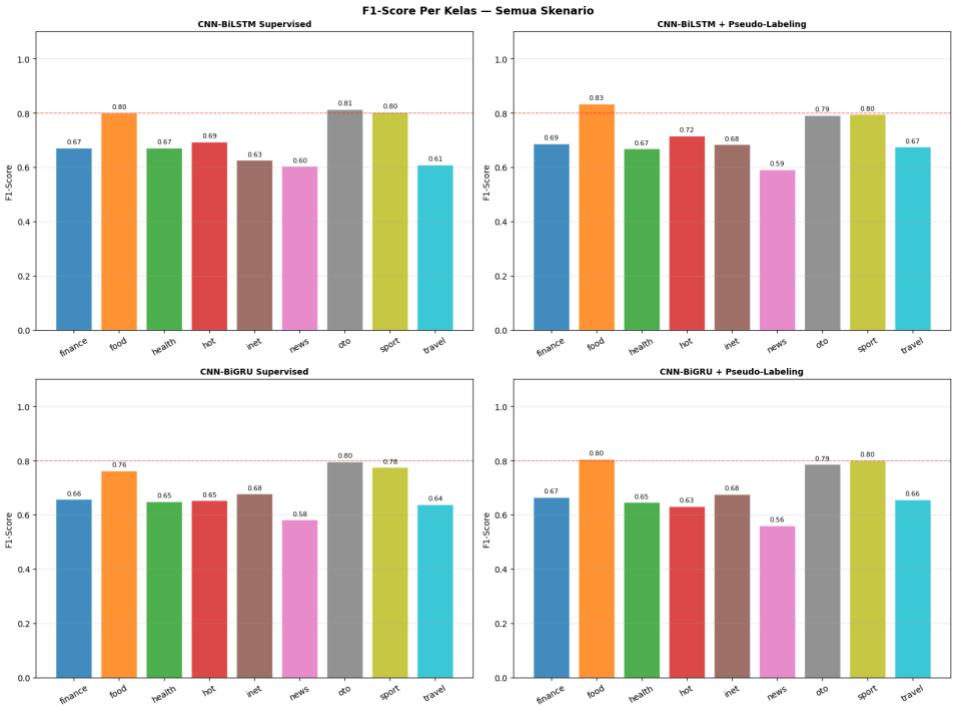


Gambar 6. Perbandingan metrik evaluasi keempat skenario model

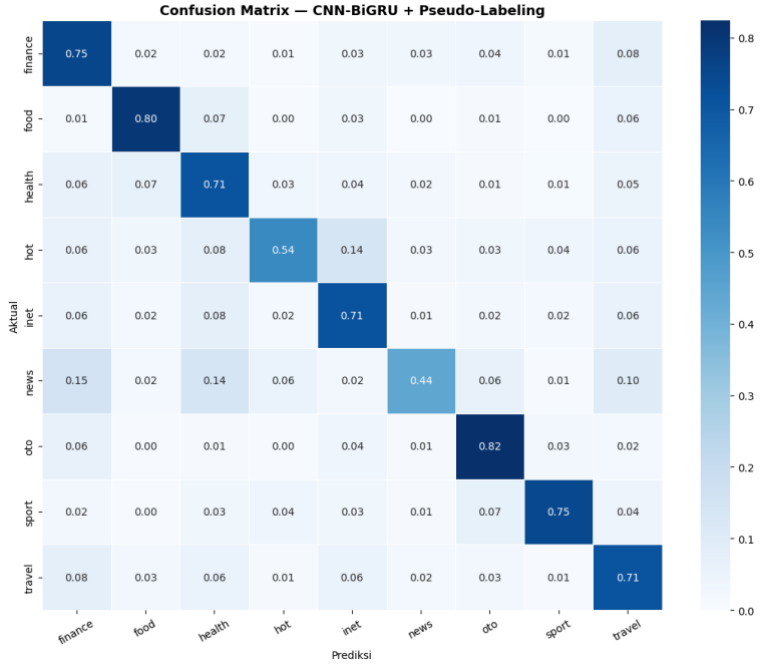


Gambar 7. Radar chart perbandingan kinerja komprehensif keempat skenario

Gambar 10 menunjukkan F1-Score per kelas untuk keempat skenario, sedangkan confusion matrix dari skenario terbaik (CNN-BiGRU + Pseudo-Labeling) ditampilkan pada Gambar 11. Dari confusion matrix terlihat bahwa kategori yang paling sulit diklasifikasikan adalah kategori yang memiliki kosakata tumpang tindih seperti *ekonomi* dan *properti*, serta *bola* dan *sport*.



Gambar 8. F1-Score per kelas untuk keempat skenario model



Gambar 9. Confusion matrix skenario terbaik (CNN-BiGRU + Pseudo-Labeling)

Secara keseluruhan, integrasi FastText *embedding* terbukti memberikan representasi teks yang kaya karena kemampuan FastText dalam menangani kata berimbuhan bahasa Indonesia melalui pendekatan *subword*. Strategi pseudo-labeling secara konsisten meningkatkan performa kedua model, karena penambahan data tak berlabel yang dipseudo-labeli membantu model mempelajari distribusi kelas yang lebih representatif [13]. Arsitektur BiGRU menunjukkan keunggulan dibandingkan BiLSTM dalam konteks ini, yang sejalan dengan temuan bahwa GRU memiliki efisiensi komputasi lebih baik dengan performa kompetitif pada teks pendek seperti judul berita [4].

4. KESIMPULAN

Penelitian ini telah membandingkan empat skenario klasifikasi judul berita Indonesia menggunakan arsitektur CNN-BiLSTM dan CNN-BiGRU yang diintegrasikan dengan FastText *embedding* bahasa Indonesia dan strategi semi-supervised berbasis *confidence-based pseudo-labeling*. Eksperimen dilakukan pada Indonesian News Dataset yang telah diseimbangkan (21.924 sampel, 9 kelas) dengan hanya 20% data berlabel. Hasil evaluasi menunjukkan bahwa CNN-BiLSTM *supervised* memperoleh akurasi 69,65%, meningkat menjadi 71,72% setelah *pseudo-labeling*, sedangkan CNN-BiGRU *supervised* memperoleh akurasi 68,62%, meningkat menjadi 69,42% setelah *pseudo-labeling*. Hasil ini menunjukkan bahwa penerapan *pseudo-labeling* secara konsisten meningkatkan performa kedua arsitektur dibandingkan pendekatan *supervised* saja, dengan CNN-BiLSTM + *pseudo-labeling* mencapai kinerja terbaik di antara keempat skenario. Hal ini mengkonfirmasi efektivitas pendekatan semi-supervised dalam kondisi keterbatasan data berlabel yang umum dalam praktik nyata. Penelitian mendatang dapat mengeksplorasi penggunaan model berbasis Transformer seperti IndoBERT sebagai perbandingan, serta penerapan teknik augmentasi data dan *curriculum pseudo-labeling* untuk lebih meningkatkan kinerja.

UCAPAN TERIMAKASIH

Penulis menyampaikan penghargaan dan terima kasih yang setinggi-tingginya kepada penyedia dan kontributor repositori data terbuka Kaggle atas ketersediaan akses dataset metadata akademik yang menjadi basis data utama dalam penelitian ini. Ketersediaan data tersebut secara terbuka turut memungkinkan terlaksananya proses eksperimen, analisis, dan pengujian sistem temu kembali informasi secara menyeluruh dalam penelitian ini. Penulis juga menyampaikan terima kasih kepada Program Studi Teknik Informatika Universitas Halu Oleo atas dukungan fasilitas dan lingkungan akademik yang mendukung kelancaran pelaksanaan penelitian ini hingga selesai.

REFERENSI

[1] Y. LeCun, Y. Bengio, dan G. Hinton, "Deep learning," Nature, vol. 521, no. 7553, hal. 436-444, 2015.
[2] C. Guridno, A. Azimah, and S. Ningsih, "Analisis Hybrid Metode CNN dan LSTM dalam Media Berita Online Indonesia," Jurnal Sistem Informasi Bisnis (JUNSIBI), vol. 5, no. 1, pp. 86–101, 2024, doi: 10.55122/junsibi.v5i1.1202.
[3] G. A. M. K. Jaluwana, G. M. A. Sasmita, and I. M. A. D. Suarjaya, "Analysis of Public Sentiment Towards Government Efforts to Break the Chain of Covid-19 Transmission in Indonesia Using CNN and Bidirectional LSTM," Jurnal RESTI: Rekayasa Sistem dan Teknologi Informasi, vol. 6, no. 4, pp. 511–520, 2022, doi: 10.29207/resti.v6i4.4055.
[4] F. Hanif, T. B. Sasongko, and A. D. Laksito, "Perbandingan Kinerja LSTM, Bi-LSTM, dan GRU pada Klasifikasi Judul Berita Clickbait," Indonesian Journal of Computer Science, vol. 12, no. 4, p. 2136, 2023.
[5] W. Widayat, "Analisis Sentimen Movie Review Menggunakan Word2Vec dan Metode LSTM Deep Learning," Jurnal Media Informasi Budidarma, vol. 5, no 3, p. 1018, 2021, doi: 10.30865/mib.v5i3.3111
[6] Nurdin et al., " Perbandingan Kinerja Word Embedding Word2Vec, GloVe, dan FastText pada Klasifikasi Teks," Jurnal Tekno Kompak, vol. 14, no. 2, pp. 74–79, 2020, doi: 10.33365/jtk.v14i2.732
[7] Yohana et al., "Pengaruh Hyperparameter pada FastText Terhadap Performa Model Deteksi Sarkasme Berbasis Bi-LSTM," JATISI: Jurnal Teknik Informatika dan Sistem Informasi, vol. 9, no. 3, pp. 2612–2624, 2022.
[8] R. Yang et al., "CNN-LSTM Deep Learning Architecture for Computer Vision-Based Modal Frequency Detection," Mechanical Systems and Signal Processing, vol. 168, 2022, doi: 10.1016/j.ymssp.2021.108657.
[9] D.-H. Lee, "Pseudo-Label: The Simple and Efficient Semi-Supervised Learning Method for Deep Neural Networks," Workshop on Challenges in Representation Learning, ICML, vol. 3, p. 896, 2013.
[10] M. D. Purbolaksono, C. Rheza, A. Munandar, dan A. Bijaksana, "Analisis sentimen pada ulasan aplikasi mobile banking menggunakan metode LSTM berbasis IndoBERT," Jurnal Informatika dan Sistem Informasi, vol. 3, no. 2, hal. 112–121, 2022.
[11] E. Grave, P. Bojanowski, P. Gupta, A. Joulin, and T. Mikolov, "Learning Word Vectors for 157 Languages," in Proc. International Conference on Language Resources and Evaluation (LREC 2018), 2018.
[12] M. N. Rizve, K. Duarte, Y. S. Rawat, and M. Shah, "In Defense of Pseudo-Labeling: An Uncertainty-Aware Pseudo-Label Selection Framework for Semi-Supervised Learning," arXiv preprint arXiv:2101.06329, 2021.
[13] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, "Bag of Tricks for Efficient Text Classification," arXiv preprint arXiv:1607.01759, 2016.
[14] M. A. Fathin, Y. Sibaroni, and S. S. Prasetyowati, "Handling Imbalance Dataset on Hoax Indonesian Political News Classification using IndoBERT and Random Sampling," Jurnal Media Informasi Budidarma, vol. 8, no. 1, p. 352, 2024, doi: 10.30865/mib.v8i1.7099.
[15] A. H. Syahlan, A. W. Hisbullah, and M. Wildan, "Klasifikasi Bahasa Menggunakan FastText," AI Dan SPK: Jurnal Artificial Intelligent Dan Sistem Penunjang Keputusan, vol. 3, no. 2, pp. 284–289, 2025.