JCOSIS (*Journal Computer Science and Information Syetem*)
E-ISSN : 3032-5676
Volume : 03 | Nomor : 02 | Tahun 2026 | Hal 116-125
DOI : https://doi.org/10.61567

Analisis Pengaruh *Label Smoothing* terhadap Ketahanan Model CNN–BiLSTM pada Klasifikasi Sentimen Ulasan Aplikasi Berbahasa Indonesia

Marissa Utami[1]*, Erwin Dwika Putra[2]
[1]Program Studi Sistem Informasi, Fakultas Teknik, Universitas Muhammadiyah Bengkulu, Indonesia
[2]Program Studi Teknik Informatika, Fakultas Teknik, Universitas Muhammadiyah Bengkulu, Indonesia
[1] marissautami@umb.ac.id ; [2] erwindwikap@gmail.com

Abstrak.
Tujuan : Penelitian ini bertujuan menganalisis pengaruh *label smoothing* terhadap ketahanan model CNN–BiLSTM dalam klasifikasi sentimen ulasan aplikasi berbahasa Indonesia pada berbagai tingkat korupsi label (*label noise*).
Metode/Design/Pendekatan: Penelitian menggunakan pendekatan eksperimen kuantitatif dengan 2.428 ulasan dari aplikasi MyTelkomsel, DANA, dan Shopee. Label sentimen ditentukan berdasarkan rating pengguna, dengan pembagian data training, validasi, dan test sebesar 70:15:15. Eksperimen melibatkan empat tingkat *label noise* (0%, 10%, 20%, dan 30%) serta empat nilai *label smoothing* (0,00; 0,05; 0,10; dan 0,20), sehingga menghasilkan 16 konfigurasi. Evaluasi dilakukan menggunakan accuracy, Macro-F1, Matthews Correlation Coefficient (MCC), dan kurva pembelajaran.
Hasil/Temuan: Nilai *label smoothing* 0,10 menghasilkan Macro-F1 sebesar 0,5831 dan MCC sebesar 0,7253 pada noise 0%, serta Macro-F1 sebesar 0,5493 dan MCC sebesar 0,6311 pada noise 20%. Namun, pada noise 10% dan 30%, konfigurasi tanpa smoothing menghasilkan Macro-F1 lebih tinggi. Kurva pembelajaran menunjukkan indikasi *overfitting*, dengan *validation loss* terendah pada epoch kedua pada konfigurasi yang dianalisis.
Kebaharuan/Originalitas/Nilai: Penelitian ini menyajikan analisis komparatif pengaruh *label smoothing* pada berbagai tingkat korupsi label dalam klasifikasi sentimen ulasan aplikasi berbahasa Indonesia. Hasil menunjukkan bahwa manfaat *label smoothing* bergantung pada tingkat *label noise* dan belum konsisten pada seluruh skenario yang diuji.

Kata Kunci: CNN–BiLSTM, klasifikasi sentimen, *label noise*, *label smoothing*, ulasan aplikasi.

Abstract.
Purpose: *This study aims to analyze the effect of label smoothing on the robustness of a CNN–BiLSTM model for Indonesian-language application review sentiment classification under different levels of label corruption (label noise).*
Methods/Study design/approach: *A quantitative experimental approach was applied to 2,428 reviews from MyTelkomsel, DANA, and Shopee. Sentiment labels were derived from user ratings, and the dataset was split into training, validation, and test sets at a ratio of 70:15:15. The experiments combined four label noise levels (0%, 10%, 20%, and 30%) with four label smoothing values (0.00, 0.05, 0.10, and 0.20), resulting in 16 configurations. Evaluation used accuracy, Macro-F1, Matthews Correlation Coefficient (MCC), and learning curves.*
Result/Findings: *A label smoothing value of 0.10 achieved a Macro-F1 of 0.5831 and an MCC of 0.7253 at 0% noise, and a Macro-F1 of 0.5493 and an MCC of 0.6311 at 20% noise. However, at noise levels of 10% and 30%, the configurations without smoothing achieved higher Macro-F1 scores. The learning curves indicated overfitting, with the lowest validation loss occurring at epoch two for the configurations examined.*
Novelty/Originality/Value: *This study provides a comparative analysis of label smoothing across different levels of label corruption in Indonesian-language application review sentiment classification. The findings indicate that the benefits of label smoothing depend on the noise level and are not consistent across all evaluated scenarios.*

Keywords*: CNN–BiLSTM, sentiment classification, label noise, label smoothing, application reviews.*

Article history:

Received, 2026-10-27
Revised, 2026-10-28
Accepted, 2026-10-31

[1]*Corresponding author.
Marissa Utami
Email addresses: marissautami@umb.ac.id

This is an open access article under the CC BY-SAlicense.

PENDAHULUAN

Perkembangan teknologi informasi dan komunikasi telah mendorong peningkatan penggunaan aplikasi seluler dalam berbagai sektor, seperti telekomunikasi, perdagangan elektronik, layanan keuangan digital, dan pelayanan publik. Peningkatan penggunaan aplikasi tersebut menghasilkan ulasan pengguna dalam jumlah besar yang memuat informasi mengenai kepuasan, keluhan, pengalaman penggunaan, dan kualitas layanan. Analisis sentimen merupakan salah satu pendekatan *Natural Language Processing* (NLP) yang dapat digunakan untuk mengidentifikasi kecenderungan opini pengguna berdasarkan teks ulasan. Informasi tersebut dapat dimanfaatkan untuk mengevaluasi kualitas aplikasi, mengidentifikasi permasalahan layanan, serta mendukung pengambilan keputusan berbasis data. Namun, klasifikasi sentimen pada ulasan aplikasi masih menghadapi tantangan berupa penggunaan bahasa informal, variasi ekspresi, ambiguitas semantik, dan ketidaksesuaian antara isi teks dengan label sentimen yang diberikan [1], [2], [3].

Dalam beberapa tahun terakhir, pendekatan deep learning telah banyak digunakan untuk meningkatkan kemampuan klasifikasi sentimen melalui pembelajaran representasi teks secara otomatis. *Convolutional Neural Network* (CNN) memiliki kemampuan untuk mengekstraksi pola lokal berupa kombinasi kata dan frasa yang mengandung informasi sentimen, sedangkan *Long Short-Term Memory* (LSTM) dapat mempelajari ketergantungan sekuensial dalam teks. Pengembangan LSTM dua arah atau *Bidirectional Long Short-Term Memory* (BiLSTM) memungkinkan pemodelan konteks dari arah maju dan mundur sehingga informasi kontekstual dapat dipertimbangkan secara lebih komprehensif. Kombinasi CNN dan BiLSTM selanjutnya memungkinkan proses ekstraksi fitur lokal dan pemodelan konteks sekuensial dilakukan dalam satu arsitektur hibrida. Sejumlah penelitian menunjukkan bahwa pendekatan CNN–LSTM dan variasinya berpotensi meningkatkan kinerja klasifikasi sentimen pada data teks yang kompleks [1], [3], [4], [5].

Meskipun memiliki kemampuan representasi yang baik, kinerja model deep learning sangat bergantung pada kualitas label data pelatihan. Dalam klasifikasi sentimen ulasan aplikasi, label sering kali ditentukan secara otomatis berdasarkan skor penilaian pengguna. Sebagai contoh, ulasan dengan rating tinggi dapat diberi label positif, sedangkan ulasan dengan rating rendah diberi label negatif [1]. Pendekatan tersebut relatif efisien untuk mengolah data dalam jumlah besar, tetapi tidak selalu mencerminkan sentimen yang sebenarnya. Ulasan dengan rating tinggi dapat mengandung keluhan tertentu, sementara rating rendah belum tentu menunjukkan bahwa seluruh isi ulasan bersifat negatif. Selain itu, sarkasme, opini campuran, dan ekspresi ambigu dapat menimbulkan ketidaksesuaian antara label yang diamati dan makna semantik teks [6].

Ketidaksesuaian label tersebut dikenal sebagai label noise, yaitu kondisi ketika label yang tersedia dalam data pelatihan tidak sepenuhnya merepresentasikan kelas yang seharusnya. Song et al. [7] menjelaskan bahwa label noise dapat menurunkan kemampuan generalisasi model dan menjadi salah satu permasalahan penting dalam pembelajaran mesin berbasis data berlabel. Model jaringan saraf yang memiliki kapasitas tinggi berpotensi mempelajari pola yang berkaitan dengan kesalahan label apabila proses pelatihan tidak mampu membedakan informasi yang relevan dari supervisi yang tidak akurat. Permasalahan ini juga telah diteliti melalui pendekatan koreksi fungsi kerugian, pengembangan fungsi kerugian yang toleran terhadap noise, dan strategi pembelajaran *robust* lainnya [7], [8], [9]. Oleh karena itu, evaluasi model klasifikasi sentimen tidak cukup hanya mempertimbangkan kinerja pada data berlabel konvensional, tetapi juga perlu mengukur ketahanannya ketika kualitas label pelatihan menurun.

Salah satu pendekatan yang berpotensi digunakan untuk mengatasi permasalahan tersebut adalah label *smoothing*. Metode ini merupakan teknik regularisasi yang mengubah distribusi target kelas yang semula bersifat tegas (*hard labels*) menjadi distribusi probabilitas yang lebih halus (*soft labels*). Pada klasifikasi multikelas, label *smoothing* memberikan sebagian probabilitas kepada kelas selain kelas target sehingga model tidak dipaksa menghasilkan probabilitas yang terlalu ekstrem terhadap satu kelas. Secara umum, pendekatan ini dapat mengurangi kecenderungan model untuk menjadi terlalu yakin terhadap prediksi selama pelatihan. Zhang et al. [10] mengkaji label smoothing secara mendalam dan mengembangkan pendekatan *Online Label Smoothing* untuk membentuk target pelatihan berdasarkan statistik prediksi model. Penelitian tersebut juga menunjukkan relevansi label smoothing terhadap peningkatan ketahanan jaringan saraf pada kondisi label yang bising. Meskipun demikian, pengaruh label smoothing perlu dievaluasi berdasarkan karakteristik tugas, distribusi kelas, dan mekanisme pembentukan label yang digunakan. Secara matematis, label *smoothing* dapat dirumuskan sebagai berikut:

$$q_k = \begin{cases} 1 - \varepsilon + \frac{\varepsilon}{K}, k = y, \\ \frac{\varepsilon}{K}, \qquad k \neq y, \end{cases} \tag{1}$$

dengan (q_k) merupakan probabilitas target setelah proses *smoothing*, (K) adalah jumlah kelas, (y) merupakan kelas target, dan (ε) adalah parameter *smoothing* dengan nilai antara 0 dan 1. Ketika $(\varepsilon = 0)$, target kembali menjadi *one-hot label konvensional*. Sementara itu, peningkatan nilai (ε) menyebabkan distribusi target menjadi lebih halus. Fungsi kerugian cross-entropy berdasarkan target tersebut dapat dituliskan sebagai:

$$\mathcal{L}LS = -\sum k = 1^{K} q_k \log p_k,$$

dengan (p_k) merupakan probabilitas prediksi model terhadap kelas ke-(k). Formulasi ini menjadi dasar untuk mengevaluasi pengaruh parameter *smoothing* terhadap proses pembelajaran CNN–BiLSTM. Namun, label smoothing tidak secara langsung mengidentifikasi atau memperbaiki label yang salah. Oleh karena itu, manfaatnya terhadap ketahanan model ketika menghadapi label *noise* harus dibuktikan melalui eksperimen, bukan diasumsikan berlaku pada seluruh kondisi data [10], [11].

Sejumlah penelitian terdahulu telah membahas penanganan label noise melalui mekanisme yang berbeda. Patrini et al. [8] mengusulkan pendekatan koreksi fungsi kerugian untuk mengurangi dampak kesalahan label yang bergantung pada kelas. Zhang dan Sabuncu [9] memperkenalkan *Generalized Cross Entropy* sebagai alternatif fungsi kerugian untuk pelatihan jaringan saraf pada data dengan label bising. Sementara itu, Song et al. [7] mengelompokkan berbagai metode pembelajaran dengan label noise dan menguraikan tantangan dalam evaluasi ketahanan model. Penelitian Zhang et al. [10] secara khusus mengkaji mekanisme label smoothing dan menunjukkan bahwa strategi pembentukan target yang lebih informatif dapat memengaruhi ketahanan model terhadap label yang tidak akurat. Temuan-temuan tersebut menunjukkan bahwa pengendalian pengaruh label noise merupakan aspek penting dalam pengembangan sistem klasifikasi yang andal. Namun, pendekatan koreksi label, fungsi kerugian *robust*, dan label *smoothing* memiliki mekanisme yang berbeda sehingga efektivitasnya tidak dapat dianggap setara tanpa pengujian komparatif.

Pada konteks analisis sentimen berbahasa Indonesia, penggunaan model hibrida CNN–BiLSTM telah menjadi salah satu pendekatan yang relevan untuk mempelajari pola lokal dan konteks sekuensial dalam ulasan pengguna. Penelitian terkait analisis sentimen bahasa Indonesia juga telah mengeksplorasi model hibrida, representasi berbasis Transformer, dan berbagai strategi pembelajaran teks [1], [3], [4], [5], [12]. Berdasarkan perkembangan tersebut, masih diperlukan pengujian yang lebih terfokus untuk memahami bagaimana tingkat label *smoothing* memengaruhi ketahanan model CNN–BiLSTM pada klasifikasi sentimen ulasan aplikasi berbahasa Indonesia [13]. Penggunaan label *smoothing* sebagai teknik regularisasi tidak dengan sendirinya menjamin peningkatan ketahanan terhadap label *noise* [14], [15]. Nilai *smoothing* yang terlalu rendah mungkin belum memberikan efek regularisasi yang berarti, sedangkan nilai yang terlalu tinggi berpotensi mengurangi ketegasan informasi kelas target. Selain itu, perubahan kinerja dapat berbeda pada setiap kelas sentimen, terutama ketika distribusi kelas tidak seimbang atau terdapat tumpang tindih semantik antara ulasan positif, negatif, dan netral [16], [17]. Oleh karena itu, penelitian perlu mengevaluasi beberapa tingkat smoothing secara sistematis dan membandingkan hasilnya dengan model dasar tanpa label smoothing. Evaluasi juga perlu membedakan kinerja pada data pengujian dengan label asli dan kinerja pada skenario pelatihan yang diberi gangguan label secara terkendali.

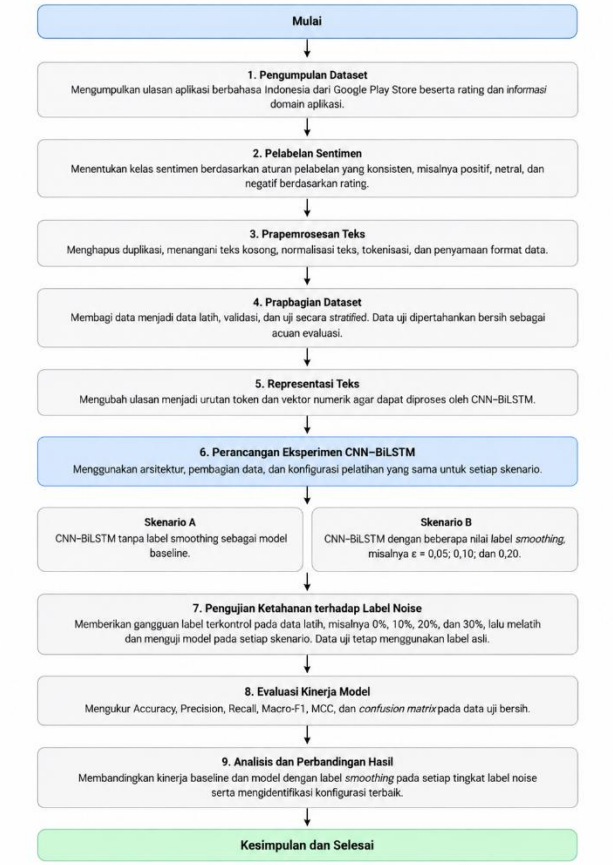
Penelitian ini bertujuan menganalisis pengaruh penerapan label *smoothing* terhadap ketahanan model CNN–BiLSTM dalam klasifikasi sentimen ulasan aplikasi berbahasa Indonesia. Eksperimen dirancang dengan membandingkan model CNN–BiLSTM tanpa label smoothing dan model dengan beberapa nilai parameter smoothing. Untuk mengukur ketahanan terhadap label noise, penelitian dapat menerapkan gangguan label sintetis dengan tingkat tertentu pada data pelatihan, sementara data validasi dan pengujian dipertahankan tanpa gangguan label buatan. Dengan rancangan tersebut, perubahan kinerja dapat diamati berdasarkan variasi nilai smoothing dan tingkat *noise* yang diberikan. Pengukuran dilakukan menggunakan akurasi, *precision*, *recall*, *Macro-F1*, dan *Matthews Correlation Coefficient* (MCC), disertai analisis confusion matrix untuk mengidentifikasi pola kesalahan pada setiap kelas. Apabila memungkinkan, eksperimen diulang menggunakan beberapa random seed untuk menilai kestabilan hasil.

Kontribusi penelitian ini mencakup tiga aspek. Pertama, memberikan evaluasi empiris mengenai pengaruh variasi parameter label smoothing terhadap kinerja model CNN–BiLSTM pada klasifikasi sentimen berbahasa Indonesia. Kedua, menganalisis perubahan kinerja model pada beberapa tingkat label noise sehingga efektivitas label *smoothing* sebagai strategi regularisasi dapat dinilai secara lebih terukur. Ketiga, mengidentifikasi kondisi ketika label *smoothing* memberikan manfaat atau justru menurunkan kinerja klasifikasi, termasuk dampaknya terhadap kelas sentimen yang memiliki karakteristik distribusi dan tingkat kesulitan berbeda. Dengan demikian, penelitian ini diharapkan dapat memberikan bukti eksperimental untuk menentukan penggunaan label smoothing yang sesuai dalam pengembangan sistem analisis sentimen ulasan aplikasi berbahasa Indonesia yang lebih *robust*.

JCOSIS (*Journal Computer Science and Information Syetem*)
E-ISSN : 3032-5676
Volume : 03 | Nomor : 02 | Tahun 2026 | Hal 116-125
DOI : https://doi.org/10.61567

METODE PENELITIAN

Alur penelitian yang ditunjukkan pada Gambar 1 menjelaskan tahapan sistematis yang dilakukan untuk menganalisis pengaruh label smoothing terhadap ketahanan model CNN–BiLSTM pada klasifikasi sentimen ulasan aplikasi berbahasa Indonesia. Penelitian ini menggunakan pendekatan eksperimen kuantitatif dengan membandingkan kinerja model pada kondisi pelatihan yang berbeda, khususnya berdasarkan penerapan label smoothing dan tingkat gangguan label (*label noise*). Tahapan penelitian meliputi pengumpulan dataset, pelabelan sentimen, prapemrosesan teks, pembagian dataset, representasi teks, perancangan model, pengujian ketahanan model, evaluasi kinerja, serta analisis hasil eksperimen. Seluruh tahapan tersebut dirancang untuk mengetahui perubahan kinerja model ketika dilatih menggunakan data yang memiliki tingkat ketidakakuratan label berbeda.



Gambar 1 Bagan alur penelitian

Tahap pertama adalah pengumpulan dataset ulasan aplikasi berbahasa Indonesia dari Google Play Store. Data yang dikumpulkan mencakup teks ulasan, rating yang diberikan pengguna, serta informasi domain aplikasi yang menjadi sumber ulasan. Dataset tersebut selanjutnya melalui proses pelabelan sentimen untuk mengelompokkan ulasan ke dalam kelas positif, netral, dan negatif berdasarkan aturan pemetaan rating yang telah ditetapkan. Penggunaan aturan pelabelan yang konsisten diperlukan agar setiap data memiliki acuan kelas yang jelas. Setelah proses pelabelan selesai, dilakukan prapemrosesan teks yang mencakup penghapusan data duplikat, penanganan teks kosong atau tidak valid, normalisasi teks, dan tokenisasi. Tahapan ini bertujuan untuk meningkatkan konsistensi data serta mempersiapkan teks ulasan agar dapat digunakan dalam proses pembelajaran model.

Tahap berikutnya adalah pembagian dataset menjadi data latih, data validasi, dan data uji menggunakan metode stratified splitting untuk mempertahankan proporsi kelas sentimen pada setiap subset. Data latih digunakan untuk membangun parameter model, data validasi digunakan untuk memantau kinerja selama pelatihan dan membantu pemilihan konfigurasi model, sedangkan data uji digunakan untuk mengevaluasi kemampuan generalisasi model setelah pelatihan selesai. Dalam penelitian ini, label pada data uji dipertahankan sesuai label aslinya dan tidak diberikan gangguan. Selanjutnya, teks ulasan dikonversi menjadi representasi numerik melalui proses tokenisasi dan embedding. Representasi tersebut digunakan sebagai masukan bagi model CNN–BiLSTM. Arsitektur CNN berfungsi mengekstraksi pola lokal dalam teks, sedangkan BiLSTM mempelajari informasi kontekstual dari

urutan teks melalui pemrosesan dua arah. Kombinasi kedua komponen tersebut digunakan untuk menghasilkan representasi yang mendukung proses klasifikasi sentimen.

Tahap eksperimen dilakukan dengan membandingkan dua skenario utama, yaitu model CNN–BiLSTM tanpa label smoothing sebagai model baseline dan model CNN–BiLSTM yang menerapkan label smoothing dengan beberapa nilai parameter, misalnya $(\varepsilon = 0{,}05), (0{,}10)$, dan $(0{,}20)$. Pada model *baseline*, proses pelatihan menggunakan fungsi kerugian dengan target label standar, sedangkan pada model eksperimen, distribusi target diperhalus menggunakan label smoothing untuk mengurangi ketergantungan model terhadap target kelas yang terlalu pasti. Untuk menguji ketahanan model, dilakukan simulasi label noise dengan tingkat gangguan 0%, 10%, 20%, dan 30% pada data latih. Gangguan diberikan secara terkontrol dengan mengubah sebagian label sesuai prosedur eksperimen yang ditetapkan. Setiap kombinasi nilai label smoothing dan tingkat gangguan diuji menggunakan arsitektur serta konfigurasi pelatihan yang konsisten agar perbedaan hasil dapat dianalisis secara lebih objektif.

Tahap terakhir adalah evaluasi dan analisis hasil eksperimen. Kinerja setiap konfigurasi model diukur menggunakan *accuracy, precision, recall, Macro-F1, Matthews Correlation Coefficient* (MCC), dan confusion matrix. Penggunaan metrik tersebut memungkinkan evaluasi kinerja klasifikasi secara menyeluruh, termasuk kemampuan model dalam mengenali setiap kelas sentimen dan menghadapi ketidakseimbangan kelas. Hasil evaluasi kemudian dibandingkan antara model baseline dan model dengan label smoothing pada setiap tingkat label noise. Analisis difokuskan pada perubahan nilai metrik ketika tingkat gangguan meningkat untuk mengetahui konfigurasi yang mampu mempertahankan kinerja klasifikasi secara lebih baik. Apabila diperlukan, setiap eksperimen dapat diulang menggunakan beberapa random seed untuk menilai konsistensi hasil. Berdasarkan hasil perbandingan tersebut, ditarik kesimpulan mengenai pengaruh label smoothing terhadap ketahanan model CNN–BiLSTM pada klasifikasi sentimen ulasan aplikasi berbahasa Indonesia.

HASIL DAN PEMBAHASAN

Penelitian ini mengevaluasi pengaruh label *smoothing* terhadap ketahanan model CNN–BiLSTM pada klasifikasi sentimen ulasan aplikasi berbahasa Indonesia. Dataset penelitian berasal dari ulasan aplikasi MyTelkomsel, DANA, dan Shopee yang mewakili domain telekomunikasi, dompet digital, dan perdagangan elektronik. Setelah proses pembersihan data, diperoleh 2.428 ulasan yang digunakan dalam eksperimen. Label sentimen ditentukan berdasarkan rating pengguna, yaitu rating 1–2 sebagai sentimen negatif, rating 3 sebagai sentimen netral, dan rating 4–5 sebagai sentimen positif. Dengan demikian, label yang digunakan merupakan label berbasis rating atau weak labels, bukan hasil anotasi sentimen secara manual. Pendekatan ini memiliki keterbatasan karena rating yang diberikan pengguna tidak selalu sepenuhnya mencerminkan sentimen yang terkandung dalam teks ulasan.

Dataset dibagi menjadi data training sebanyak 1.700 ulasan, data validasi sebanyak 365 ulasan, dan data test sebanyak 363 ulasan. Pembagian data dilakukan dengan mempertimbangkan kelompok teks untuk menghindari kebocoran teks identik yang telah dinormalisasi antarsubset. Data training terdiri atas 801 ulasan negatif, 79 ulasan netral, dan 820 ulasan positif. Distribusi tersebut memperlihatkan bahwa kelas netral memiliki jumlah sampel yang jauh lebih sedikit dibandingkan kelas negatif dan positif. Kondisi ini perlu diperhatikan dalam interpretasi hasil evaluasi, terutama pada metrik *Macro-F1* yang memberikan bobot yang sama terhadap setiap kelas.

Eksperimen dilakukan dengan mengombinasikan empat tingkat label noise, yaitu 0%, 10%, 20%, dan 30%, dengan empat nilai label smoothing, yaitu 0,00; 0,05; 0,10; dan 0,20, sehingga menghasilkan 16 kombinasi eksperimen. Label noise diterapkan secara artifisial pada label data training dengan mengubah sebagian label ke salah satu kelas lainnya. Label validasi dan test tidak diberi noise tambahan dan tetap menggunakan label awal yang berasal dari rating. Dengan demikian, eksperimen ini menguji ketahanan model terhadap korupsi label training yang dikendalikan secara artifisial, bukan secara langsung mengukur seluruh bentuk kesalahan anotasi sentimen yang mungkin terjadi pada data nyata. Model CNN–BiLSTM dilatih menggunakan optimizer Adam, ukuran *batch* 32, *learning rate* 0,001, dan batas maksimum 30 atau 50 epoch. Proses pelatihan menggunakan *EarlyStopping* berdasarkan *validation loss* dengan *patience* 5 serta mekanisme pengurangan learning rate ketika performa validasi tidak membaik. Evaluasi dilakukan menggunakan *accuracy, Macro-Precision, Macro-Recall, Macro-F1, Weighted-F1, Matthews Correlation Coefficient* (MCC), dan *validation loss* terbaik.

JCOSIS (*Journal Computer Science and Information Syetem*)
E-ISSN : 3032-5676
Volume : 03 | Nomor : 02 | Tahun 2026 | Hal 116-125
DOI : https://doi.org/10.61567

Tabel 1. Hasil pengujian 16 kombinasi label noise dan label smoothing

Label noise	Label smoothing	Accuracy	Macro-Precision	Macro-Recall	Macro-F1	Weighted-F1	MCC	Validation loss
0%	0,00	0,8430	0,5648	0,5899	0,5757	0,8233	0,7063	0,5986
0%	0,05	0,8512	0,5692	0,5956	0,5813	0,8312	0,7208	0,6455
0%	0,10	0,8540	0,5704	0,5974	0,5831	0,8338	0,7253	0,6909
0%	0,20	0,8512	0,5688	0,5955	0,5812	0,8311	0,7203	0,7791
10%	0,00	0,8402	0,5610	0,5878	0,5737	0,8203	0,6986	0,5836
10%	0,05	0,8320	0,5554	0,5820	0,5680	0,8123	0,6825	0,6451
10%	0,10	0,8320	0,5570	0,5821	0,5681	0,8123	0,6848	0,7148
10%	0,20	0,8347	0,5586	0,5840	0,5700	0,8150	0,6897	0,8009
20%	0,00	0,7906	0,5306	0,5534	0,5398	0,7719	0,6071	0,6701
20%	0,05	0,7989	0,5354	0,5591	0,5455	0,7800	0,6218	0,7106
20%	0,10	0,8044	0,5381	0,5628	0,5493	0,7854	0,6311	0,7469
20%	0,20	0,8017	0,5369	0,5610	0,5474	0,7827	0,6268	0,8294
30%	0,00	0,8264	0,5510	0,5777	0,5636	0,8060	0,6720	0,6782
30%	0,05	0,8182	0,5461	0,5718	0,5576	0,7975	0,6573	0,7166
30%	0,10	0,8182	0,5464	0,5717	0,5575	0,7974	0,6580	0,7622
30%	0,20	0,8182	0,5458	0,5718	0,5578	0,7977	0,6568	0,8370

Berdasarkan Tabel 1, performa model bervariasi menurut tingkat label noise dan nilai label smoothing. Pada kondisi tanpa label noise, konfigurasi smoothing 0,10 memperoleh accuracy sebesar 0,8540 dan Macro-F1 sebesar 0,5831. Pada tingkat noise 10%, konfigurasi tanpa smoothing menghasilkan Macro-F1 sebesar 0,5737, lebih tinggi dibandingkan konfigurasi dengan *smoothing* 0,05, 0,10, dan 0,20. Pada tingkat *noise* 20%, smoothing 0,10 menghasilkan Macro-F1 sebesar 0,5493, sedangkan pada tingkat noise 30%, baseline tanpa smoothing memperoleh Macro-F1 sebesar 0,5636. Hasil ini menunjukkan bahwa peningkatan nilai smoothing tidak selalu diikuti peningkatan performa klasifikasi. Oleh karena itu, pengaruh label smoothing perlu dianalisis pada setiap tingkat noise, bukan hanya berdasarkan satu konfigurasi dengan nilai metrik tertinggi. Nilai validation loss pada tabel digunakan untuk menggambarkan perilaku model selama pelatihan, sedangkan metrik test digunakan untuk melaporkan performa pada data evaluasi.

Untuk mengidentifikasi perubahan performa akibat penerapan label smoothing, setiap konfigurasi dibandingkan dengan baseline tanpa smoothing pada tingkat noise yang sama. Perubahan Macro-F1 dihitung menggunakan persamaan ($\Delta F1_{\text{macro}} = F1_{\text{macro,smoothing}} - F1_{\text{macro,baseline}}$). Nilai positif menunjukkan peningkatan Macro-F1 terhadap baseline, sedangkan nilai negatif menunjukkan penurunan. Hasil perhitungan disajikan pada Tabel 2.

Tabel 2. Perubahan Macro-F1 terhadap baseline tanpa smoothing

Label noise	Baseline smoothing 0,00	Perubahan pada 0,05	Perubahan pada 0,10	Perubahan pada 0,20
0%	0,5757	+0,0055	+0,0074	+0,0055
10%	0,5737	−0,0057	−0,0056	−0,0037
20%	0,5398	+0,0057	+0,0094	+0,0076
30%	0,5636	−0,0059	−0,0061	−0,0058

Hasil pada Tabel 2 memperlihatkan bahwa pada tingkat noise 0%, seluruh nilai *smoothing* menghasilkan peningkatan Macro-F1 dibandingkan *baseline*. Peningkatan terbesar terjadi pada *smoothing* 0,10, yaitu sebesar 0,0074. Pada tingkat noise 10%, seluruh konfigurasi smoothing menghasilkan penurunan Macro-F1, dengan penurunan berkisar antara 0,0037 dan 0,0057. Pada tingkat *noise* 20%, seluruh konfigurasi *smoothing* kembali menunjukkan peningkatan Macro-F1, dengan peningkatan terbesar sebesar 0,0094 pada *smoothing* 0,10. Sebaliknya, pada tingkat noise 30%, seluruh konfigurasi smoothing menghasilkan penurunan sekitar 0,0058–0,0061. Dengan demikian, smoothing 0,10 memberikan hasil yang lebih baik daripada baseline pada dua kondisi, yaitu *noise* 0% dan 20%, tetapi tidak pada *noise* 10% dan 30%. Pola ini memperlihatkan bahwa pengaruh *smoothing* tidak bersifat konsisten pada seluruh tingkat korupsi label. Karena setiap konfigurasi hanya dijalankan satu kali, perbedaan tersebut diperlakukan sebagai temuan deskriptif dan belum dapat dinyatakan sebagai peningkatan yang stabil atau signifikan secara statistik.

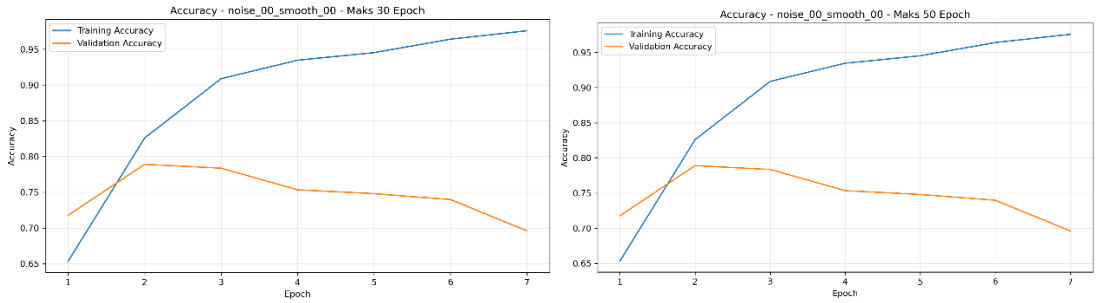
Selain Macro-F1, analisis dilakukan menggunakan *Matthews Correlation Coefficient* (MCC) untuk memperoleh gambaran tambahan mengenai kesesuaian hasil klasifikasi. Perubahan MCC dihitung dengan membandingkan nilai MCC setiap konfigurasi *smoothing* terhadap baseline pada tingkat *noise* yang sama, menggunakan persamaan ($\Delta MCC = MCC_{\text{smoothing}} - MCC_{\text{baseline}}$). Hasilnya disajikan pada Tabel 3.

Tabel 3. Perubahan MCC terhadap *baseline* tanpa *smoothing*

Label noise	Baseline smoothing 0,00	Perubahan pada 0,05	Perubahan pada 0,10	Perubahan pada 0,20
0%	0,7063	+0,0145	+0,0189	+0,0139
10%	0,6986	−0,0161	−0,0139	−0,0089
20%	0,6071	+0,0147	+0,0240	+0,0197
30%	0,6720	−0,0147	−0,0140	−0,0152

Berdasarkan Tabel 3, perubahan MCC memperlihatkan pola yang sejalan dengan hasil Macro-F1. Pada noise 0%, seluruh nilai smoothing meningkatkan MCC dibandingkan baseline, dengan peningkatan terbesar sebesar 0,0189 pada smoothing 0,10. Pada noise 20%, smoothing 0,10 juga memberikan peningkatan terbesar, yaitu 0,0240. Sebaliknya, pada noise 10% dan 30%, seluruh konfigurasi smoothing menghasilkan MCC yang lebih rendah daripada baseline. Kesesuaian pola antara Macro-F1 dan MCC memperkuat temuan deskriptif bahwa manfaat smoothing bergantung pada skenario eksperimen. Meskipun demikian, hasil ini belum membuktikan bahwa smoothing secara umum lebih efektif pada tingkat noise tertentu. Nilai MCC baseline pada noise 30%, misalnya, lebih tinggi daripada baseline pada noise 20%. Hal ini menunjukkan bahwa performa tidak berubah secara monoton mengikuti persentase noise. Perbedaan tersebut dapat dipengaruhi oleh sampel yang terkena perubahan label, distribusi kelas, dan variasi proses pelatihan.

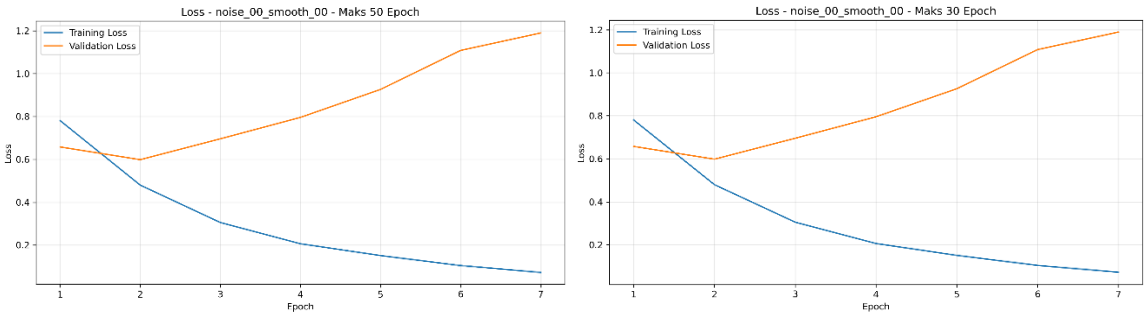
Analisis kurva pembelajaran digunakan untuk memahami perilaku model selama training dan validasi. Pada konfigurasi baseline dengan noise 0% dan smoothing 0,00, training accuracy meningkat dari sekitar 65,3% pada epoch pertama menjadi sekitar 97,6% pada epoch ketujuh. Sebaliknya, validation accuracy mencapai nilai tertinggi sekitar 78,9% pada epoch kedua, kemudian menurun hingga sekitar 69,6% pada epoch ketujuh. Perbedaan pola tersebut menunjukkan bahwa peningkatan kemampuan model dalam memprediksi data training tidak diikuti peningkatan performa pada data validasi. Kondisi ini merupakan indikasi bahwa model mulai mengalami kesulitan melakukan generalisasi terhadap data yang tidak digunakan dalam pelatihan.



Gambar 2. Kurva accuracy baseline

Kurva loss pada konfigurasi baseline memperkuat temuan tersebut. Training loss menurun dari sekitar 0,78 pada epoch pertama menjadi sekitar 0,07 pada epoch ketujuh, sedangkan validation loss mencapai nilai minimum sekitar 0,60 pada epoch kedua dan kemudian meningkat hingga sekitar 1,19 pada epoch ketujuh. Penurunan training loss yang berlangsung bersamaan dengan peningkatan validation loss merupakan indikasi kuat terjadinya overfitting. Model semakin menyesuaikan bobotnya terhadap data training, tetapi performa pada data validasi justru memburuk. Dengan demikian, penambahan jumlah maksimum epoch tidak selalu meningkatkan kualitas model ketika performa validasi telah mengalami penurunan.

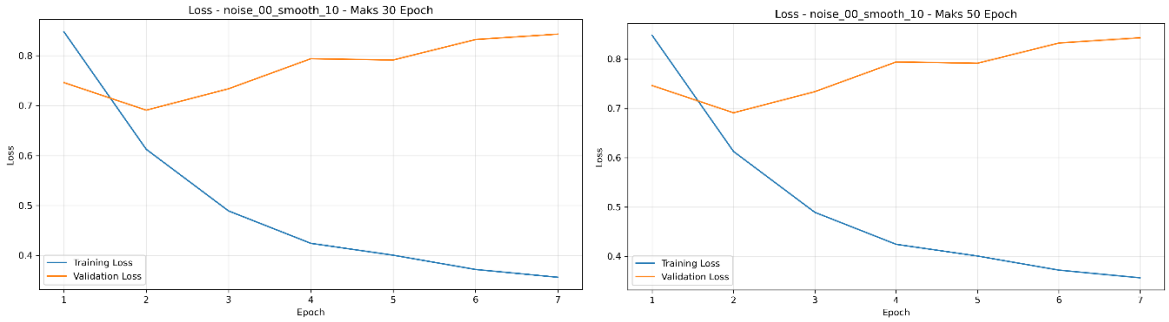
Pada konfigurasi noise 0% dan smoothing 0,10, training loss juga terus menurun, sedangkan validation loss mencapai nilai minimum sekitar 0,69 pada epoch kedua dan meningkat hingga sekitar 0,84 pada epoch ketujuh. Pola ini menunjukkan bahwa overfitting masih terlihat meskipun smoothing diterapkan. Oleh karena itu, grafik yang tersedia belum memberikan bukti bahwa smoothing 0,10 menghilangkan overfitting. Namun, perbandingan langsung nilai loss antarkonfigurasi perlu dilakukan secara hati-hati karena smoothing mengubah distribusi target yang digunakan dalam perhitungan loss. Evaluasi manfaat smoothing perlu mempertimbangkan metrik validasi yang sama, seperti Macro-F1, MCC, dan accuracy, selain kurva loss.



Gambar 3. Kurva loss baseline

Hasil eksperimen dengan batas maksimum 30 dan 50 epoch menunjukkan nilai metrik yang identik. Berdasarkan catatan pelatihan, seluruh konfigurasi mencapai epoch terbaik pada epoch kedua dan berhenti setelah tujuh epoch dijalankan. Hal ini menunjukkan bahwa batas maksimum epoch tidak menentukan jumlah epoch aktual ketika EarlyStopping menghentikan pelatihan lebih awal karena validation loss tidak membaik. Untuk memastikan evaluasi menggunakan model terbaik, mekanisme pemulihan bobot pada epoch terbaik perlu diperiksa, misalnya melalui pengaturan restore_best_weights=True pada EarlyStopping atau penggunaan checkpoint terbaik. Dengan demikian, evaluasi akhir tidak menggunakan bobot epoch terakhir apabila performa validasi terbaik diperoleh pada epoch sebelumnya.

Secara keseluruhan, hasil eksperimen menunjukkan bahwa pengaruh label smoothing terhadap ketahanan model CNN–BiLSTM bergantung pada tingkat label noise yang diterapkan. Smoothing 0,10 menunjukkan peningkatan Macro-F1 dan MCC terhadap baseline pada noise 0% dan 20%, tetapi menghasilkan penurunan pada noise 10% dan 30%. Hasil ini memperlihatkan bahwa label smoothing tidak secara otomatis meningkatkan ketahanan model pada seluruh kondisi korupsi label. Selain itu, kurva pembelajaran menunjukkan indikasi overfitting pada konfigurasi baseline dan smoothing 0,10 yang dianalisis. Temuan tersebut memberikan gambaran awal mengenai hubungan antara label smoothing, korupsi label training, dan generalisasi model, tetapi belum cukup untuk menetapkan nilai smoothing yang optimal secara umum.



Gambar 4. Kurva loss dengan label smoothing 0,10

Penelitian ini memiliki beberapa keterbatasan yang perlu dipertimbangkan dalam menginterpretasikan hasil. Pertama, setiap kombinasi eksperimen hanya dijalankan satu kali sehingga hasil belum menggambarkan variasi akibat pengacakan inisialisasi bobot, urutan batch, maupun pemilihan label yang diubah. Kedua, label sentimen awal ditentukan berdasarkan rating pengguna, bukan anotasi manual terhadap isi ulasan, sehingga kemungkinan ketidaksesuaian antara rating dan sentimen tekstual tetap ada. Ketiga, jumlah sampel kelas netral pada data test hanya 17 ulasan, sehingga beberapa perubahan prediksi dapat memengaruhi nilai metrik kelas dan Macro-F1. Keempat, kesimpulan mengenai pengaruh smoothing perlu dibatasi pada dataset, arsitektur, parameter, dan skenario noise yang diuji. Eksperimen dengan beberapa random seed dapat dilakukan pada penelitian lanjutan untuk mengevaluasi kestabilan hasil dan mengetahui apakah pola yang ditemukan tetap muncul pada pengulangan pelatihan.

Berdasarkan hasil yang diperoleh, dapat disimpulkan bahwa label smoothing menunjukkan pengaruh yang tidak seragam terhadap performa CNN–BiLSTM pada klasifikasi sentimen ulasan aplikasi berbahasa Indonesia. Dalam eksperimen ini, smoothing 0,10 memberikan peningkatan Macro-F1 dan MCC pada tingkat noise 0% dan 20%, tetapi tidak pada tingkat noise 10% dan 30%. Oleh karena itu, hasil penelitian belum menunjukkan bahwa label smoothing meningkatkan ketahanan model secara konsisten. Kesimpulan ini dibatasi pada hasil satu kali pelatihan

untuk setiap konfigurasi dan tidak dimaksudkan sebagai bukti bahwa nilai smoothing tertentu merupakan nilai optimal secara umum.

KESIMPULAN

Penelitian ini menunjukkan bahwa pengaruh label smoothing terhadap ketahanan model CNN–BiLSTM pada klasifikasi sentimen ulasan aplikasi berbahasa Indonesia bergantung pada tingkat korupsi label training. Dari 16 konfigurasi yang diuji, nilai *smoothing* 0,10 menghasilkan Macro-F1 sebesar 0,5831 dan MCC sebesar 0,7253 pada tingkat *noise* 0%, serta Macro-F1 sebesar 0,5493 dan MCC sebesar 0,6311 pada tingkat noise 20%. Sebaliknya, pada tingkat noise 10% dan 30%, konfigurasi tanpa *smoothing* menghasilkan Macro-F1 yang lebih tinggi daripada konfigurasi dengan *smoothing*. Analisis kurva pembelajaran juga memperlihatkan indikasi overfitting, ditandai oleh peningkatan *validation loss* setelah epoch kedua pada konfigurasi yang diamati. Dengan demikian, label smoothing belum terbukti meningkatkan ketahanan model secara konsisten pada seluruh tingkat noise. Kesimpulan ini dibatasi pada hasil satu kali pelatihan untuk setiap konfigurasi, sehingga penelitian selanjutnya perlu menguji kestabilan hasil melalui pengulangan eksperimen dan mengevaluasi performa berdasarkan metrik validasi serta test yang digunakan secara tepat.

REFERENSI

[1] V. R. Prasetyo, M. F. Naufal, and K. Wijaya, "Sentiment Analysis of ChatGPT on Indonesian Text using Hybrid CNN and Bi-LSTM," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 9, no. 2, pp. 327–333, Apr. 2025, doi: 10.29207/resti.v9i2.6334.

[2] H. Murfi, S. Theresia Gowandi, G. Ardaneswari, and S. Nurrohmah, "BERT-based combination of convolutional and recurrent neural network for indonesian sentiment analysis," *Appl. Soft Comput.*, vol. 151, 2024, doi: 10.1016/j.asoc.2023.111112.

[3] P. K. Jain, V. Saravanan, and R. Pamula, "A Hybrid CNN-LSTM: A Deep Learning Approach for Consumer Sentiment Analysis Using Qualitative User-Generated Contents," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, vol. 20, no. 5, pp. 1–15, Sep. 2021, doi: 10.1145/3457206.

[4] H. Murfi, Syamsyuriani, T. Gowandi, G. Ardaneswari, and S. Nurrohmah, "BERT-based combination of convolutional and recurrent neural network for indonesian sentiment analysis," *Appl. Soft Comput.*, vol. 151, p. 111112, Jan. 2024, doi: 10.1016/j.asoc.2023.111112.

[5] M. Umer, I. Ashraf, A. Mehmood, S. Kumari, S. Ullah, and G. Sang Choi, "Sentiment analysis of tweets using a unified convolutional neural network-long short-term memory network model," *Comput. Intell.*, vol. 37, no. 1, pp. 409–434, 2021, doi: 10.1111/coin.12415.

[6] M. Atabuzzaman, M. Shajalal, M. B. Baby, and A. Boden, "Arabic Sentiment Analysis with Noisy Deep Explainable Model," *ACM Int. Conf. Proceeding Ser.*, pp. 185–189, 2023, doi: 10.1145/3639233.3639241.

[7] H. Song, M. Kim, D. Park, Y. Shin, and J.-G. Lee, "Learning From Noisy Labels With Deep Neural Networks: A Survey," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 34, no. 11, pp. 8135–8153, Nov. 2023, doi: 10.1109/TNNLS.2022.3152527.

[8] G. Patrini, A. Rozza, A. K. Menon, R. Nock, and L. Qu, "Making Deep Neural Networks Robust to Label Noise: A Loss Correction Approach," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Jul. 2017, pp. 2233–2241. doi: 10.1109/CVPR.2017.240.

[9] Z. Zhang and M. R. Sabuncu, "Generalized Cross Entropy Loss for Training Deep Neural Networks with Noisy Labels," Nov. 2018, [Online]. Available: http://arxiv.org/abs/1805.07836

[10] C.-B. Zhang *et al.*, "Delving Deep Into Label Smoothing," *IEEE Trans. Image Process.*, vol. 30, pp. 5984–5996, 2021, doi: 10.1109/TIP.2021.3089942.

[11] H. Park, J. Noh, Y. Oh, D. Baek, and B. Ham, "ACLS: Adaptive and Conditional Label Smoothing for Network Calibration," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, IEEE, Oct. 2023, pp. 3913–3922. doi: 10.1109/ICCV51070.2023.00364.

[12] H. D. Abubakar and M. Umar, "Sentiment Classification: Review of Text Vectorization Methods: Bag of Words, Tf-Idf, Word2vec and Doc2vec," *SLU J. Sci. Technol.*, vol. 4, no. 1&2, pp. 27–33, Aug. 2022, doi: 10.56471/slujst.v4i.266.

[13] W. Meng, Y. Wei, P. Liu, Z. Zhu, and H. Yin, "Aspect Based Sentiment Analysis With Feature Enhanced Attention CNN-BiLSTM," *IEEE Access*, vol. 7, pp. 167240–167249, 2019, doi: 10.1109/ACCESS.2019.2952888.

[14] P. K. Jain, V. Saravanan, and R. Pamula, "A Hybrid CNN-LSTM: A Deep Learning Approach for Consumer Sentiment Analysis Using Qualitative User-Generated Contents," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, vol. 20, no. 5, 2021, doi: 10.1145/3457206.

[15] J. Fu, H. Zheng, and T. Mei, "Look Closer to See Better: Recurrent Attention Convolutional Neural Network for Fine-Grained Image Recognition," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Jul. 2017, pp. 4476–4484. doi: 10.1109/CVPR.2017.476.

[16] Y. Fu and Y. Liu, "Cross-domain sentiment classification based on key pivot and non-pivot extraction," *Knowledge-Based Syst.*, vol. 228, no. 2, p. 107280, Sep. 2021, doi: 10.1016/j.knosys.2021.107280.

[17] D. Liu, I. W. Tsang, and G. Yang, "A Convergence Path to Deep Learning on Noisy Labels," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 35, no. 4, pp. 5170–5182, Apr. 2024, doi: 10.1109/TNNLS.2022.3202752.