



Identifikasi Komplain di Lingkungan Perguruan Tinggi Menggunakan Algorithm Convolutional Long Short Term Memory dan FastText Embedding

Received : Agus 06, 2026; Revised : Agus 17, 2026; Accepted : Sep 03, 2026; Published : Sep 15, 2026

Khusnul Khotimah^{1*}, Cahyo Crysdian², M.Imamudin³

¹Universitas Islam Negeri Maulana Mallik Ibrahim,
Jl Gajayana No. 50, Dinoyo, Lowokwaru, Malang, Indonesia
khusnul23z4@gmail.com

² Universitas Islam Negeri Maulana Mallik Ibrahim,
Jl Gajayana No. 50, Dinoyo, Lowokwaru, Malang, Indonesia
cahyo@ti.uin-malang.ac.id

³Universitas Islam Negeri Maulana Mallik Ibrahim,
Jl Gajayana No. 50, Dinoyo, Lowokwaru, Malang, Indonesia
imamudin@ti.uin-malang.ac.id

*Corresponding Author

Abstrak:

Perguruan tinggi rutin menerima keluhan mahasiswa terkait akademik, dosen, tenaga kependidikan, pembelajaran, fasilitas, dan keuangan, yang jika diklasifikasikan secara manual memakan waktu dan rawan subjektivitas. Penelitian ini mengusulkan sistem klasifikasi keluhan otomatis berbasis Long Short-Term Memory (LSTM) yang dipadukan dengan representasi kata FastText. Sebanyak 835 komentar keluhan dari lima program studi dikumpulkan dan dibersihkan menjadi 449 komentar unik, dikelompokkan ke dalam enam kategori: akademik, keuangan, layanan dosen, layanan tenaga kependidikan, pembelajaran, dan fasilitas. Setelah case folding, tokenisasi, dan padding, teks direpresentasikan dengan embedding FastText berdimensi 300 yang dilatih khusus pada data latih tiap fold untuk mencegah kebocoran data. Delapan arsitektur berbasis LSTM (LSTM, Bi-LSTM, Attention-LSTM, Attention-Bi-LSTM, serta varian gabungannya dengan lapisan konvolusi) dikembangkan dan dibandingkan, dengan class weighting untuk mengatasi ketidakseimbangan data. Evaluasi dilakukan melalui 5x5 repeated cross-validation (25 pengulangan) dan data uji terpisah (80:20) menggunakan metrik akurasi, presisi, recall, dan F1-score. Model Convolutional Attention-LSTM (CNN-A-LSTM) menunjukkan kinerja paling konsisten, dengan akurasi data uji 71,11% dan macro F1-score 0,719 (rata-rata validasi silang $0,732 \pm 0,037$), mengungguli model dasar LSTM (akurasi 66,67%, F1-score 0,687; validasi silang $0,655 \pm 0,047$). Paired t-test pada 25 pasangan validasi silang membuktikan ketujuh varian arsitektur meningkat signifikan dibanding model dasar ($p < 0,05$), bahkan setelah koreksi Holm-Bonferroni dan Benjamini-Hochberg. Temuan ini menegaskan bahwa penambahan konvolusi, konteks dua arah, dan atensi pada representasi FastText secara konsisten meningkatkan klasifikasi keluhan mahasiswa, mendukung pengelolaan keluhan yang lebih cepat, objektif, dan efisien di perguruan tinggi.

Kata Kunci :

Klasifikasi Keluhan, Convolutional LSTM, Pembelajaran Mendalam (Deep Learning), FastText, Perguruan Tinggi, Klasifikasi Teks.

Abstract:

Higher education institutions routinely receive student complaints related to academics, lecturers, administrative staff, learning processes, facilities, and finance, which, when classified manually, are



time-consuming and prone to subjectivity. This study proposes an automatic complaint classification system based on Long Short-Term Memory (LSTM) combined with FastText word representations. A total of 835 student complaint comments from five study programs were collected and cleaned into 449 unique comments, grouped into six categories: academic, finance, lecturer services, administrative staff services, learning, and facilities. After case folding, tokenization, and padding, the texts were represented using 300-dimensional FastText embeddings trained exclusively on each fold's training data to prevent data leakage. Eight LSTM-based architectures (LSTM, Bi-LSTM, Attention-LSTM, Attention-Bi-LSTM, and their respective variants combined with convolutional layers) were developed and compared, with class weighting applied to address data imbalance. Evaluation was carried out using 5x5 repeated cross-validation (25 repetitions) and a separate test set (80:20 split), using accuracy, precision, recall, and F1-score as metrics. The Convolutional Attention-LSTM (CNN-A-LSTM) model showed the most consistent performance, achieving a test-set accuracy of 71.11% and a macro F1-score of 0.719 (cross-validation mean of 0.732 ± 0.037), outperforming the baseline LSTM model (accuracy of 66.67%, F1-score of 0.687; cross-validation mean of 0.655 ± 0.047). A paired t-test on 25 cross-validation pairs demonstrated that all seven architectural variants improved significantly over the baseline model ($p < 0.05$), even after Holm-Bonferroni and Benjamini-Hochberg multiple-comparison correction. These findings confirm that adding convolution, bidirectional context, and attention mechanisms to FastText representations consistently improves student complaint classification, supporting faster, more objective, and more efficient complaint management in higher education institutions.

Keywords:

Complaint Classification, Convolutional LSTM, Deep Learning, FastText, Higher Education, Text Classification.

1. Pendahuluan

Perguruan tinggi dituntut untuk terus meningkatkan mutu layanan pendidikan sebagai bagian dari penjaminan mutu internal dan akreditasi. Salah satu sumber informasi penting untuk perbaikan mutu adalah umpan balik (feedback) mahasiswa yang berupa keluhan atau komplain terhadap berbagai aspek layanan, mulai dari layanan akademik, kinerja dosen, tenaga kependidikan, proses pembelajaran, hingga sarana prasarana dan keuangan. Umpan balik berbentuk teks bebas mengandung informasi yang kaya dan rinci, namun analisisnya secara manual membutuhkan waktu lama dan cenderung subjektif [1], [2].

Seiring meningkatnya jumlah komplain yang masuk melalui kanal e-complaint, proses membaca, mengelompokkan, dan mendistribusikan komplain ke unit terkait menjadi beban tersendiri bagi institusi. Oleh karena itu, diperlukan sistem klasifikasi teks otomatis yang mampu mengidentifikasi kategori komplain secara cepat dan konsisten. Klasifikasi teks merupakan salah satu tugas mendasar dalam pemrosesan bahasa alami (natural language processing/NLP) yang telah banyak dikaji, khususnya pada analisis sentimen dan opini [3], [4].

Dalam beberapa tahun terakhir, pendekatan pembelajaran mendalam (deep learning) terbukti unggul dibandingkan metode pembelajaran mesin klasik untuk klasifikasi teks. Model berbasis Long Short-Term Memory (LSTM) dan variannya, seperti Bidirectional LSTM (Bi-LSTM), mampu menangkap ketergantungan konteks antar kata pada kalimat, sedangkan Convolutional Neural Network (CNN) efektif mengekstraksi fitur lokal berupa frasa atau n-gram [5], [6], [7], [8]. Penggabungan CNN dan LSTM menghasilkan arsitektur hibrida Convolutional LSTM yang memanfaatkan kekuatan keduanya secara sekaligus [9], [10], [11].

Kualitas representasi teks (word embedding) sangat menentukan performa klasifikasi. FastText, yang dikembangkan dengan memanfaatkan informasi subword (n-gram karakter), lebih tangguh

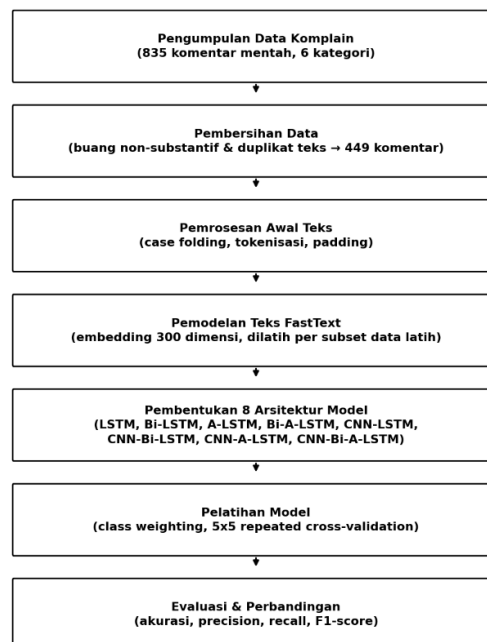
terhadap variasi penulisan, imbuhan, dan kata di luar kosakata (out-of-vocabulary) dibandingkan Word2Vec dan GloVe [12]. Karakteristik ini sangat relevan untuk komplain mahasiswa yang umumnya ditulis secara informal dan tidak baku. Beberapa penelitian menunjukkan bahwa kombinasi FastText dengan arsitektur CNN-BiLSTM dan mekanisme atensi memberikan hasil yang sangat baik pada berbagai bahasa [13], [14], [15]

Meskipun demikian, sebagian besar penelitian klasifikasi teks berfokus pada analisis sentimen produk atau media sosial, dan masih terbatas kajian yang secara khusus mengidentifikasi komplain di lingkungan perguruan tinggi menggunakan data riil berbahasa Indonesia yang tidak seimbang antar kelas. Selain itu, perbandingan menyeluruh antar delapan arsitektur berbasis LSTM (dengan dan tanpa konvolusi maupun atensi) pada kasus ini belum banyak dilakukan.

Berdasarkan latar belakang tersebut, penelitian ini bertujuan membangun dan membandingkan delapan arsitektur berbasis LSTM yang dipadukan dengan embedding FastText untuk mengidentifikasi komplain mahasiswa perguruan tinggi ke dalam enam kategori. Kontribusi penelitian ini adalah: (1) penerapan algoritma Convolutional LSTM dan FastText pada data komplain riil perguruan tinggi berbahasa Indonesia; (2) perbandingan menyeluruh delapan arsitektur menggunakan akurasi, precision, recall, dan F1-score; serta (3) penanganan ketidakseimbangan kelas melalui pembobotan kelas dan pengujian signifikansi statistik hasil perbandingan..

2. Metode

Penelitian ini menggunakan pendekatan eksperimental dengan tahapan sebagaimana ditunjukkan pada Gambar 1, yaitu pengumpulan data, pemrosesan awal teks, pemodelan teks dengan FastText, pembentukan model, pelatihan, serta evaluasi dan perbandingan model.



Gambar 1: Alur tahapan penelitian

2.1 Dataset

Data yang digunakan merupakan komplain riil mahasiswa Universitas Merdeka Pasuruan dari lima program studi (Teknik Informatika, Rekayasa Perangkat Lunak, Manajemen, Agroteknologi, dan Ilmu

Hukum) yang dikumpulkan melalui kuesioner, dimana kategori komplain telah ditetapkan sejak awal pada instrumen kuesioner. Kuesioner disusun dalam enam bagian pertanyaan terbuka yang masing-masing mewakili satu kategori akademik, keuangan, layanan dosen, layanan tenaga kependidikan, pembelajaran, dan sarana prasarana sehingga setiap komentar secara otomatis terkait dengan satu label sesuai bagian tempat komentar diisi (pelabelan bersifat tunggal / single-label by design, tanpa anotator manual tambahan). Definisi operasional tiap kategori adalah akademik (kurikulum, jadwal kuliah, nilai, dan administrasi akademik), keuangan (pembayaran, biaya kuliah, dan beasiswa), layanan dosen (kualitas mengajar, kehadiran, dan responsivitas dosen), layanan tenaga kependidikan (pelayanan administrasi dan staf non-dosen), pembelajaran (metode, materi, dan proses belajar-mengajar), serta sarana prasarana (fasilitas fisik seperti ruang kelas, laboratorium, toilet, dan jaringan internet). dari kuisoner tersebut menghasilkan 835 baris data mentah. Proses pembersihan dilakukan dalam dua tahap yaitu membuang 34 komentar yang tidak substantif atau tidak menyatakan keluhan (mis. "tidak ada", "baik", "-", "cukup"), dan menghapus duplikasi teks yang identik persis (373 baris), sehingga diperoleh 449 komentar unik yang layak diklasifikasikan ke dalam enam kategori komplain, yaitu akademik, keuangan, layanan dosen, layanan tenaga kependidikan (tendik), pembelajaran, dan sarana prasarana (sarpras). Distribusi data bersifat tidak seimbang sebagaimana disajikan pada Tabel 1, dengan kelas sarpras mendominasi sekitar seperempat data.

Tabel 1: Distribusi kelas komplain pada dataset

Kelas Komplain	Jumlah	Persentase
Sarpras	105	23,4%
Akademik	77	17,1%
Layanan Dosen	72	16,0%
Pembelajaran	72	16,0%
Keuangan	62	13,8%
Layanan Tendik	61	13,6%
Total	449	100,0%

2.2 Pemrosesan Awal Teks

Tahap pemrosesan awal teks meliputi case folding (mengubah seluruh huruf menjadi huruf kecil), pembersihan karakter non-alfabet (angka dan tanda baca), tokenisasi (memecah kalimat menjadi token kata), serta padding sehingga setiap komentar memiliki panjang seragam 30 token. Pemilihan metode pemrosesan awal yang tepat berpengaruh signifikan terhadap performa klasifikasi teks [16].

2.3 Pemodelan Teks dengan FastText

Setiap token direpresentasikan menjadi vektor berdimensi 300 menggunakan FastText. Untuk mencegah kebocoran data, model FastText dilatih ulang secara terpisah pada setiap subset data latih sesuai skema evaluasi (pada data latih tiap fold saat validasi silang, dan pada 80% data latih saat evaluasi akhir), bukan pada seluruh korpus sekaligus. Pada data latih akhir (359 komentar) diperoleh 1.057 kata unik. FastText dipilih karena kemampuannya membentuk vektor kata berbasis subword sehingga mampu menangani kata tidak baku dan kesalahan ketik yang lazim pada komplain [12],[13]. Matriks embedding yang dihasilkan digunakan sebagai bobot lapisan embedding yang dibekukan (non-trainable) agar perbandingan antar arsitektur bersifat adil.

2.4 Arsitektur Model

Delapan arsitektur berbasis LSTM dibangun di atas lapisan embedding FastText. LSTM mengendalikan aliran informasi melalui gerbang lupa, gerbang masukan, dan gerbang keluaran; sebagai contoh, gerbang lupa dirumuskan pada Persamaan (1), dengan σ sebagai fungsi sigmoid, W bobot, dan b bias [17].

$$f_t = \sigma(W^f \cdot [h_{t-1}, x_t] + b^f) \quad (1)$$

Bi-LSTM memproses urutan token dari dua arah (maju dan mundur), mekanisme atensi memberi bobot lebih besar pada token yang paling menentukan kelas, sedangkan lapisan konvolusi dan pooling berperan sebagai ekstraktor fitur lokal sebelum LSTM. Kedelapan kombinasi tersebut adalah LSTM, Bi-LSTM, A-LSTM (attention-LSTM), Bi-A-LSTM, CNN-LSTM, CNN-Bi-LSTM, CNN-A-LSTM, dan CNN-Bi-A-LSTM. Lapisan keluaran menggunakan fungsi softmax pada Persamaan (2) untuk menghasilkan probabilitas enam kelas.

$$\hat{y} = \text{softmax}(W \cdot h + b) \quad (2)$$

Untuk reproduktibilitas, rincian implementasi setiap arsitektur adalah sebagai berikut. Sel LSTM menggunakan fungsi sigmoid pada gerbang lupa, masukan, dan keluaran, serta fungsi tanh pada kandidat sel dan keluaran tersembunyi (konfigurasi baku LSTM). Lapisan konvolusi menggunakan fungsi aktivasi ReLU. Mekanisme atensi yang digunakan pada varian A-LSTM/Bi-A-LSTM berbentuk atensi aditif (Bahdanau) skor dihitung sebagai $\tanh(W \cdot h + b)$, bobot atensi diperoleh melalui softmax atas skor tersebut, kemudian vektor konteks dibentuk sebagai jumlah berbobot dari seluruh keluaran LSTM pada tiap langkah waktu. Dropout (0,5) diterapkan pada vektor fitur akhir sebelum lapisan keluaran. Bobot embedding FastText dibekukan (tidak dilatih ulang) selama pelatihan classifier. Seluruh eksperimen menggunakan random seed tetap (lihat Tabel 2) untuk inisialisasi bobot dan pengacakan batch, agar hasil dapat direproduksi. Bobot kelas (class weighting balanced) dihitung berbanding terbalik dengan frekuensi tiap kelas pada data latih.

Konfigurasi hyperparameter yang diterapkan secara seragam pada seluruh model dirangkum pada Tabel 2. Nilai-nilai tersebut diadopsi dari praktik umum pada literatur klasifikasi teks berbasis CNN-BiLSTM dan FastText yang sejenis (bukan hasil pencarian grid/random search formal, mengingat keterbatasan ukuran dataset), dan digunakan secara seragam pada seluruh delapan arsitektur agar perbandingan antar-model tetap adil.

Tabel 2: Konfigurasi hyperparameter pelatihan model

Parameter	Nilai
Dimensi embedding (FastText)	300
Panjang input (max_len)	30 token
Unit LSTM	64
Lapisan konvolusi	2 lapis (64 filter, kernel 3) + max pooling
Dropout	0,5
Optimizer	Adam (learning rate = 0,001)
Fungsi loss	Categorical Cross-Entropy
Batch size	32
Epoch maksimum	40 (early stopping, patience = 7)
Penanganan imbalance	Class weighting (balanced)
Skema validasi	5x5 repeated cross-validation (25 pengulangan)
Random seed	42 (basis; bervariasi sistematis per pengulangan)
Pembagian data	80% latih : 20% uji (stratified)

2.5 Pelatihan dan Evaluasi

Data dibagi menjadi 80% data latih dan 20% data uji secara stratified. Untuk mengatasi ketidakseimbangan kelas, diterapkan pembobotan kelas sehingga kelas minoritas memperoleh bobot lebih besar [18], [19], [20]. Setiap model dievaluasi menggunakan 5x5 repeated cross-validation (5-fold diulang sebanyak 5 kali dengan pembagian fold yang berbeda tiap pengulangan, menghasilkan 25 titik data per model) pada data latih, dan pengujian akhir pada data uji. Metrik evaluasi yang digunakan adalah akurasi, precision, recall, dan F1-score sebagaimana dirumuskan pada Persamaan (3)–(6), dengan TP, TN, FP, dan FN berturut-turut menyatakan true positive, true negative, false positive, dan false negative. F1-score makro menjadi acuan utama karena memberi bobot setara pada seluruh kelas termasuk kelas minoritas.

$$Precision = \frac{TP}{TP+FP} \quad (3)$$

$$Recall = \frac{TP}{TP+FN} \quad (4)$$

$$F1 - Score = 2 \times \frac{precision \times recall}{precision+recall} \quad (5)$$

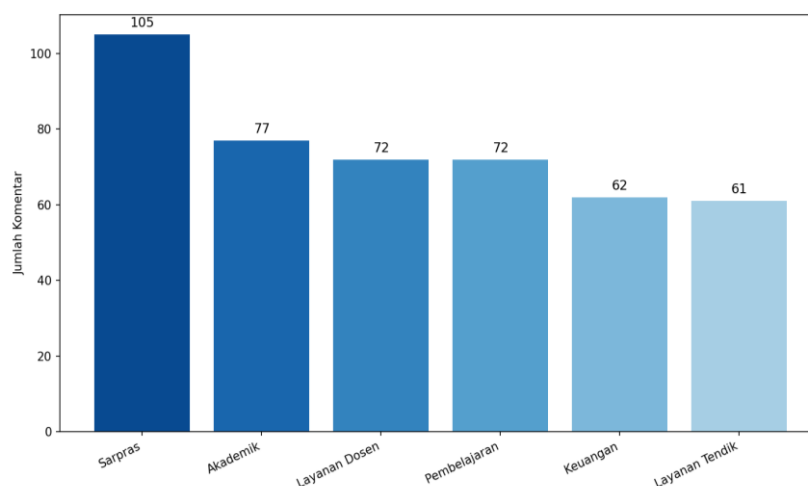
$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (6)$$

Untuk menilai signifikansi perbedaan performa antar model terhadap baseline LSTM, dilakukan uji paired t-test terhadap nilai F1-score makro pada seluruh 25 titik data 5x5 repeated cross-validation dengan taraf signifikansi 5%, dilengkapi effect size (Cohen's d) serta koreksi multiple-comparison Holm-Bonferroni dan Benjamini-Hochberg mengingat terdapat tujuh perbandingan sekaligus terhadap baseline.

3. Hasil dan Pembahasan

3.1 Karakteristik Dataset

Distribusi jumlah komentar untuk setiap kategori komplain ditampilkan pada Gambar 2. Terlihat bahwa kelas sarpras jauh lebih dominan dibandingkan kelas layanan dosen dan layanan tendik, sehingga penanganan ketidakseimbangan kelas menjadi penting agar kelas minoritas tidak terabaikan.



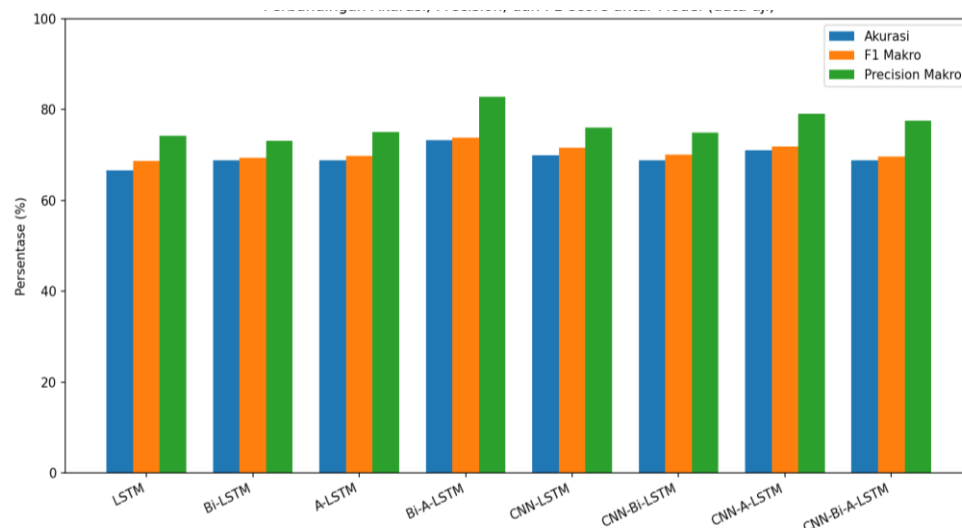
Gambar 2: Distribusi jumlah komentar per kelas komplain

3.2 Perbandingan Performa Model

Hasil pengujian kedelapan model pada data uji ($n=90$) dirangkum pada Tabel 3 dan divisualisasikan pada Gambar 3. Pada data uji tunggal ini, model Bi-A-LSTM mencatat F1-score makro tertinggi (0,738), namun mengingat kecilnya ukuran data uji, evaluasi yang lebih andal diperoleh dari rata-rata 5x5 repeated cross-validation (kolom F1-Score CV). Berdasarkan rata-rata tersebut, model Convolutional Attention-LSTM (CNN-A-LSTM) memperoleh performa paling konsisten (F1 makro CV $0,732 \pm 0,037$; akurasi data uji 71,11%, F1 makro data uji 0,719), sedangkan model baseline LSTM memperoleh performa terendah pada kedua skema evaluasi (akurasi data uji 66,67%, F1 makro data uji 0,687; F1 makro CV $0,655 \pm 0,047$). Hal ini menegaskan bahwa penambahan mekanisme atensi dan/atau konvolusi pada representasi FastText secara konsisten meningkatkan kemampuan klasifikasi dibandingkan LSTM polos, sebagaimana dikonfirmasi lebih lanjut melalui uji signifikansi statistik pada Bagian 3.4.

Tabel 3: Perbandingan performa kedelapan model pada data uji

Model	Akurasi	Precision	Recall	F1-Score	F1-Score CV
LSTM	66,67%	0,743	0,659	0,687	0,655
Bi-LSTM	68,89%	0,731	0,684	0,694	0,708
A-LSTM	68,89%	0,750	0,681	0,697	0,731
Bi-A-LSTM	73,33%	0,828	0,719	0,738	0,718
CNN-LSTM	70,00%	0,761	0,702	0,715	0,708
CNN-Bi-LSTM	68,89%	0,749	0,690	0,701	0,715
CNN-A-LSTM	71,11%	0,791	0,693	0,719	0,732
CNN-Bi-A-LSTM	68,89%	0,775	0,681	0,697	0,707



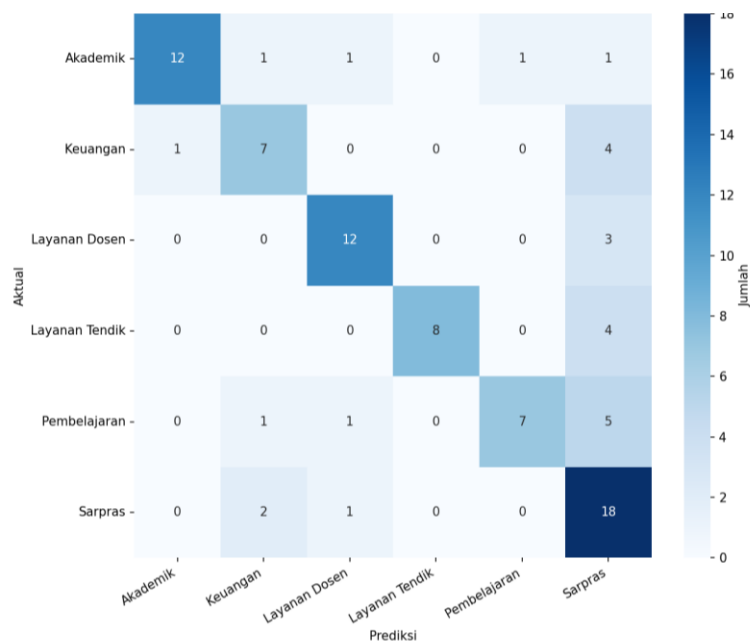
Gambar 3: Perbandingan akurasi dan F1-score antar model

Menariknya, model paling kompleks (CNN-Bi-A-LSTM) tidak mencatat performa terbaik meski jauh mengungguli baseline. Berdasarkan rata-rata validasi silang, dua model dengan mekanisme atensi namun struktur lebih sederhana, yaitu CNN-A-LSTM dan A-LSTM, justru menempati peringkat teratas, sementara penggabungan atensi, konvolusi, dan pemrosesan dua arah sekaligus (CNN-Bi-A-LSTM) tidak memberikan tambahan manfaat. Hal ini mengindikasikan bahwa pada dataset berukuran kecil, penambahan kompleksitas arsitektur secara berlebihan tidak selalu berbanding lurus dengan performa, sedangkan mekanisme atensi secara tersendiri memberikan kontribusi yang cukup konsisten. Temuan

ini sejalan dengan penelitian yang menunjukkan bahwa efektivitas arsitektur hibrida bergantung pada ukuran dan karakteristik data [21], [22], [23].

3.3 Analisis Model Terbaik

Confusion matrix model CNN-A-LSTM pada data uji ditunjukkan pada Gambar 4, sedangkan rincian performa per kelas disajikan pada Tabel 4. Model bekerja paling baik pada kelas akademik (F1-score 0,828), diikuti layanan dosen dan layanan tendik (F1-score 0,800 pada keduanya) yang cenderung memiliki penanda leksikal khas. Sebaliknya, performa terendah terjadi pada kelas keuangan (F1-score 0,609) dan sarpras (F1-score 0,643), sarpras justru menjadi kelas dengan precision terendah (0,514) meskipun jumlah datanya terbanyak, menandakan model cenderung salah mengarahkan sejumlah komentar dari kelas lain (terutama keuangan) ke kelas sarpras. Pola ini konsisten dengan tumpang tindih makna antar kelas yang memang lazim terjadi pada data komplain bebas-teks.



Gambar 4: Confusion matrix model CNN-A-LSTM pada data uji

Tabel 4: Laporan klasifikasi per kelas model CNN-A-LSTM

Kelas	Precision	Recall	F1-Score	Support
Akademik	0,923	0,750	0,828	16
Keuangan	0,636	0,583	0,609	12
Layanan Dosen	0,800	0,800	0,800	15
Layanan Tendik	1,000	0,667	0,800	12
Pembelajaran	0,875	0,500	0,636	14
Sarpras	0,514	0,857	0,643	21
Akurasi			0,711	90
Rata-rata Makro	0,792	0,693	0,719	90
Rata-rata Tertimbang	0,772	0,711	0,717	90

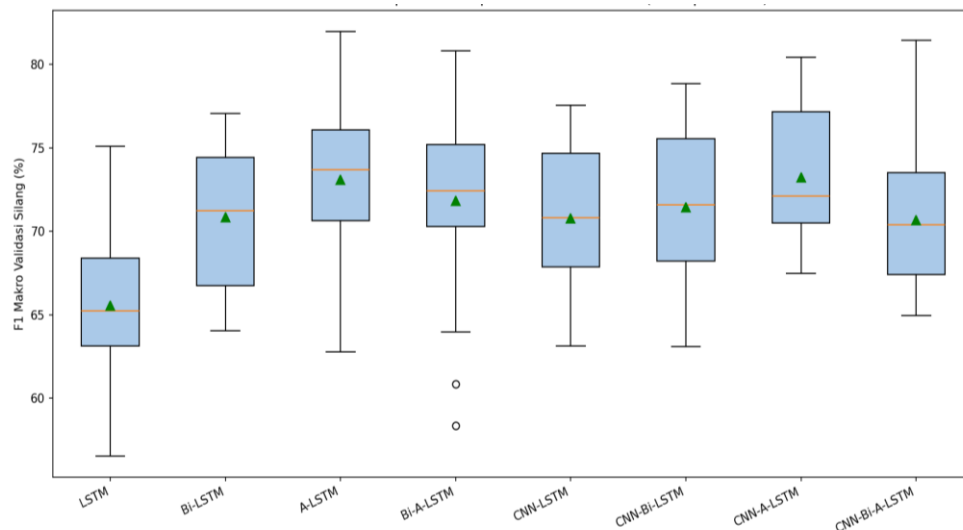
3.4 Uji Signifikansi Statistik

Untuk memastikan bahwa peningkatan performa terhadap baseline LSTM bukan sekadar perbedaan acak, dilakukan uji paired t-test terhadap nilai F1-score makro pada 5x5 repeated cross-validation (25 pasangan per model). Ukuran sampel ini sengaja diperbesar dari skema 5-fold tunggal karena $n=5$ dinilai terlalu kecil untuk memberikan daya uji (power) yang memadai serta rentan terhadap kesalahan akibat pengujian berganda (multiple comparisons) tanpa koreksi. Hasilnya disajikan pada Tabel 5, lengkap dengan effect size (Cohen's d) dan koreksi multiple-comparison menggunakan prosedur Holm-Bonferroni dan Benjamini-Hochberg (FDR). Seluruh tujuh model varian memperoleh peningkatan yang signifikan secara statistik ($p < 0,05$) dibandingkan baseline, dan signifikansi ini tetap bertahan setelah kedua jenis koreksi multiple-comparison diterapkan. Hal ini menegaskan bahwa penambahan mekanisme bidirectional, atensi, maupun konvolusi memberikan kontribusi nyata terhadap performa klasifikasi komplain, dengan tingkat keyakinan statistik yang jauh lebih kuat dibandingkan pengujian pada sampel kecil.

Tabel 5: Hasil paired t-test F1-score makro (5x5 repeated cross-validation, $n=25$) terhadap baseline LSTM

Model (vs LSTM)	Nilai t	Nilai p	Cohen's d	Signifikan (Holm-Bonferroni)	Signifikan (Benjamini-Hochberg)
Bi-LSTM	6,363	$1,40 \times 10^{-6}$	1,273	Ya	Ya
A-LSTM	7,187	$1,99 \times 10^{-7}$	1,437	Ya	Ya
Bi-A-LSTM	4,875	$5,71 \times 10^{-5}$	0,975	Ya	Ya
CNN-LSTM	4,900	$5,35 \times 10^{-5}$	0,980	Ya	Ya
CNN-Bi-LSTM	6,168	$2,26 \times 10^{-6}$	1,234	Ya	Ya
CNN-A-LSTM	8,373	$1,39 \times 10^{-8}$	1,675	Ya	Ya
CNN-Bi-A-LSTM	6,356	$1,43 \times 10^{-6}$	1,271	Ya	Ya

Distribusi F1-score makro pada 5x5 repeated cross-validation (Gambar 5) memperlihatkan bahwa model dengan mekanisme atensi (A-LSTM dan CNN-A-LSTM) cenderung memiliki median lebih tinggi dan sebaran yang relatif stabil dibandingkan varian lain. Secara keseluruhan, hasil penelitian menunjukkan bahwa kombinasi FastText dengan Convolutional Attention-LSTM merupakan pendekatan yang paling konsisten untuk mengidentifikasi komplain mahasiswa pada dataset ini. Ekstraksi fitur lokal oleh konvolusi membantu menangkap frasa penanda kategori, sementara mekanisme atensi membantu model berfokus pada bagian kalimat yang paling relevan dengan kategori komplain. Hasil ini konsisten dengan berbagai penelitian yang melaporkan keunggulan arsitektur hibrida CNN-BiLSTM berbasis embedding pada klasifikasi teks [13], [11], [24], [25], [26]. Untuk memperkuat keandalan statistik, evaluasi pada penelitian ini telah menggunakan 5x5 repeated cross-validation (25 pengulangan). Meskipun demikian, keterbatasan utama penelitian ini adalah ukuran data yang tetap relatif kecil (449 komentar) dan tidak seimbang serta berasal dari satu perguruan tinggi, sehingga performa pada kelas minoritas (keuangan, layanan tendik) masih perlu ditingkatkan dan generalisasi ke perguruan tinggi lain perlu diuji lebih lanjut.



Gambar 5: Distribusi F1-score makro pada 5x5 repeated cross-validation (n=25)

4. Kesimpulan

Penelitian ini membangun dan membandingkan delapan arsitektur berbasis LSTM yang dipadukan dengan embedding FastText untuk mengidentifikasi komplain mahasiswa perguruan tinggi ke dalam enam kategori. Berdasarkan 449 data komplain riil hasil pembersihan dari 835 data mentah, model Convolutional Attention-LSTM (CNN-A-LSTM) memperoleh performa paling konsisten dengan akurasi data uji 71,11%, precision makro 0,791, recall makro 0,693, dan F1-score makro 0,719 (rata-rata 5x5 repeated cross-validation $0,732 \pm 0,037$), mengungguli model baseline LSTM (akurasi data uji 66,67%; F1-score makro 0,687; rata-rata validasi silang $0,655 \pm 0,047$). Uji paired t-test pada 25 pasangan validasi silang menunjukkan bahwa peningkatan performa seluruh tujuh varian arsitektur terhadap baseline bersifat signifikan secara statistik ($p < 0,05$) dan tetap bertahan setelah koreksi multiple-comparison Holm-Bonferroni maupun Benjamini-Hochberg, sehingga penambahan konvolusi, pemrosesan dua arah, dan khususnya mekanisme atensi terbukti berkontribusi nyata terhadap performa klasifikasi.

Dengan mempertimbangkan keseimbangan antara performa dan kompleksitas, model CNN-A-LSTM direkomendasikan untuk diterapkan pada sistem identifikasi komplain di lingkungan perguruan tinggi, mengingat performanya paling konsisten pada evaluasi berulang dan tidak memerlukan komponen bidirectional yang menambah biaya komputasi tanpa manfaat sepadan pada dataset ini. Penelitian selanjutnya disarankan menambah jumlah data khususnya pada kelas minoritas (keuangan dan layanan tendik), menerapkan teknik augmentasi teks atau oversampling, menguji generalisasi model pada data dari perguruan tinggi lain, serta mengeksplorasi representasi berbasis transformer seperti IndoBERT sebagai pembanding untuk menangani komplain yang panjang dan bermakna ganda.

Pustaka

- [1] Z. Kastrati, F. Dalipi, A. S. Imran, K. P. Nuci, and M. A. Wani, "Sentiment Analysis of Students' Feedback with NLP and Deep Learning: A Systematic Mapping Study," *Applied Sciences* 2021, Vol. 11, Page 3986, vol. 11, no. 9, p. 3986, Apr. 2021, doi: 10.3390/APP11093986.
- [2] A. Koufakou, "Deep learning for opinion mining and topic classification of course reviews," *Education and Information Technologies* 2023 29:3, vol. 29, no. 3, pp. 2973–2997, Jun. 2023, doi: 10.1007/S10639-023-11736-2.

- [3] M. Birjali, M. Kasri, and A. Beni-Hssane, "A comprehensive survey on sentiment analysis: Approaches, challenges and trends," *Knowl. Based. Syst.*, vol. 226, p. 107134, Aug. 2021, doi: 10.1016/J.KNOSYS.2021.107134.
- [4] E. M. Mercha and H. Benbrahim, "Machine learning and deep learning for sentiment analysis across languages: A survey," *Neurocomputing*, vol. 531, pp. 195–216, Apr. 2023, doi: 10.1016/J.NEUCOM.2023.02.015.
- [5] S. Susandri, S. Defit, and M. Tajuddin, "Enhancing Text Sentiment Classification with Hybrid CNN-BiLSTM Model on WhatsApp Group", doi: 10.12720/jait.15.3.355-363.
- [6] C. H. Lin and U. Nuha, "Sentiment analysis of Indonesian datasets based on a hybrid deep-learning strategy," *Journal of Big Data 2023 10:1*, vol. 10, no. 1, pp. 88–, May 2023, doi: 10.1186/S40537-023-00782-9.
- [7] D. G. Mandhasiya, H. Murfi, and A. Bustamam, "The hybrid of BERT and deep learning models for Indonesian sentiment analysis," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 33, no. 1, pp. 591–602, Jan. 2024, doi: 10.11591/IJEECS.V33.I1.PP591-602.
- [8] G. Z. Nabillah, I. N. Alam, E. S. Purwanto, and M. F. Hidayat, "Indonesian multilabel classification using IndoBERT embedding and MBERT classification," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 14, no. 1, pp. 1071–1078, Feb. 2024, doi: 10.11591/ijece.v14i1.pp1071-1078.
- [9] Z. Anwar, H. Afzal, A. Ahsan, N. Iltaf, and A. Maqbool, "A novel hybrid CNN-LSTM approach for assessing StackOverflow post quality," *Journal of Intelligent Systems*, vol. 32, no. 1, Jan. 2023, doi: 10.1515/JISYS-2023-0057/XML.
- [10] Z. L. Zhai, X. Zhang, F. F. Fang, and L. Y. Yao, "Text classification of Chinese news based on multi-scale CNN and LSTM hybrid model," *Multimed. Tools Appl.*, vol. 82, no. 14, pp. 20975–20988, Jun. 2023, doi: 10.1007/S11042-023-14450-W.
- [11] K. Kaur and P. Kaur, "Improving BERT model for requirements classification by bidirectional LSTM-CNN deep model," *Computers and Electrical Engineering*, vol. 108, p. 108699, May 2023, doi: 10.1016/J.COMPELECENG.2023.108699.
- [12] E. M. Dharma, F. Lumban Gaol, H. L. H. S. Warnars, and B. Soewito, "The accuracy comparison among word2vec, glove, and fasttext towards convolution neural network (CNN) text classification," *Journal of Theoretical and Applied Information Technology*, vol. 100, no. 2, pp. 349–359, Jan. 2022.
- [13] S. Lee, Y. Yang, I. Aiyanyo, M. Keith, M. Boussougou, and D.-J. Park, "Attention-Based 1D CNN-BiLSTM Hybrid Model Enhanced with FastText Word Embedding for Korean Voice Phishing Detection," *Mathematics 2023, Vol. 11, Page 3217*, vol. 11, no. 14, p. 3217, Jul. 2023, doi: 10.3390/MATH11143217.
- [14] Y. Liu *et al.*, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," Jul. 2019, Accessed: Aug. 06, 2026. [Online]. Available: <http://arxiv.org/abs/1907.11692>
- [15] A. HE and M. ABISADO, "Text Sentiment Analysis of Douban Film Short Comments Based on BERT-CNN-BiLSTM-Att Model," *IEEE*, vol. 12, Mar. 2024, Accessed: Aug. 06, 2026. [Online]. Available: <https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=10478509>

- [16] C. P. Chai, "Comparison of text preprocessing methods," *Nat. Lang. Eng.*, vol. 29, no. 3, pp. 509–553, May 2023, doi: 10.1017/S1351324922000213.
- [17] I. D. Mienye, T. G. Swart, and G. Obaido, "Recurrent Neural Networks: A Comprehensive Review of Architectures, Variants, and Applications," *Information 2024, Vol. 15, Page 517*, vol. 15, no. 9, p. 517, Aug. 2024, doi: 10.3390/INFO15090517.
- [18] S. F. Abdhood, N. Omar, and S. Tiun, "Data augmentation for Arabic text classification: a review of current methods, challenges and prospective directions," *PeerJ Comput. Sci.*, vol. 11, 2025, doi: 10.7717/PEERJ-CS.2685/.
- [19] A. A. Khan, O. Chaudhari, and R. Chandra, "A review of ensemble learning and data augmentation models for class imbalanced problems: Combination, implementation and evaluation," *Expert Syst. Appl.*, vol. 244, Jun. 2024, doi: 10.1016/J.ESWA.2023.122778;PAGE:STRING:ARTICLE/CHAPTER.
- [20] M. Arslan and C. Cruz, "Business text classification with imbalanced data and moderately large label spaces for digital transformation," *Applied Network Science 2024 9:1*, vol. 9, no. 1, pp. 11–, Apr. 2024, doi: 10.1007/S41109-024-00623-5.
- [21] J. Mutinda, W. Mwangi, and G. Okeyo, "Sentiment Analysis of Text Reviews Using Lexicon-Enhanced Bert Embedding (LeBERT) Model with Convolutional Neural Network," *Applied Sciences 2023, Vol. 13, Page 1445*, vol. 13, no. 3, p. 1445, Jan. 2023, doi: 10.3390/APP13031445.
- [22] J. Wang, Z. Li, and C. Ma, "Research on News Text Classification Based on BERT-BiLSTM-TextCNN-Attention," *ACM International Conference Proceeding Series*, pp. 295–298, Mar. 2024, doi: 10.1145/3672919.3672973.
- [23] M. R. R. Rana, A. Nawaz, T. Ali, A. M. El-Sherbeeney, and W. Ali, "A BiLSTM-CF and BiGRU-based Deep Sentiment Analysis Model to Explore Customer Reviews for Effective Recommendations," *Engineering, Technology & Applied Science Research*, vol. 13, no. 5, pp. 11739–11746, Oct. 2023, doi: 10.48084/ETASR.6278.
- [24] K. Saputra, K. E. Saputra, and R. Riccosan, "Indonesian news article authorship attribution multilabel multiclass classification using IndoBERT," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 13, no. 4, pp. 4688–4694, Dec. 2024, doi: 10.11591/ijai.v13.i4.pp4688-4694.
- [25] O. Karakaya and Z. H. Kilimci, "An efficient consolidation of word embedding and deep learning techniques for classifying anticancer peptides: FastText+BiLSTM," *PeerJ Comput. Sci.*, vol. 10, p. e1831, 2024, doi: 10.7717/PEERJ-CS.1831/SUPP-8.
- [26] T. Sabri, S. Bahassine, O. El Beggat, and M. Kissi, "An improved Arabic text classification method using word embedding," *International Journal of Electrical and Computer Engineering*, vol. 14, no. 1, pp. 721–731, Feb. 2024, doi: 10.11591/IJECE.V14I1.PP721-731.