 ISSN:2302-8149
e-ISSN:2540-9719

Sistem Otomatis Deteksi dan Perbaikan Kesopanan Teks Akademik Berbasis *IndoBERT* dan *IndoT5*

An Automated System for Detecting and Improving Academic Text Politeness Using IndoBERT and IndoT5

[1]Belina Eka Ariesty, [2]Chanifah Indah Ratnasari*
[1,2]Program Studi Informatika, Fakultas Teknologi Industri, Universitas Islam Indonesia
[1,2]Jl. Kaliurang km 14.5, Sleman, Yogyakarta 55584, Indonesia
*e-mail: chanifah.indah@uii.ac.id

(***received*****:** 29 January 2026, ***revised*****:** 25 February 2026, ***accepted*****:** 27 February 2026)

Abstrak

Meningkatnya penggunaan komunikasi digital dalam interaksi akademik antara mahasiswa dan dosen sering kali tidak diiringi dengan penerapan norma kesopanan berbahasa yang konsisten, sehingga berpotensi memengaruhi efektivitas interaksi akademik. Hingga saat ini, upaya peningkatan kesopanan berbahasa umumnya masih dilakukan secara manual dan bersifat subjektif. Penelitian ini bertujuan mengembangkan sebuah sistem otomatis untuk mendeteksi dan memperbaiki kesopanan teks komunikasi akademik berbahasa Indonesia. Metode yang digunakan berupa integrasi model *IndoBERT* sebagai pengklasifikasi tingkat kesopanan teks dan *IndoT5* sebagai model generatif untuk memperbaiki kalimat yang terdeteksi tidak sopan. *Dataset* penelitian berjumlah 6.230 kalimat yang dikumpulkan melalui GoogleForm, TikTok, serta data sintetis menggunakan *ChatGPT* yang telah diberi label. Hasil evaluasi menunjukkan bahwa model *IndoBERT* mencapai akurasi sebesar 97,11% dalam mengklasifikasikan kesopanan teks akademik, sementara model *IndoT5* mampu memperbaiki kalimat tidak sopan menjadi bentuk yang lebih sesuai dengan kaidah komunikasi akademik, sebagaimana ditunjukkan oleh hasil evaluasi menggunakan metrik *BLEU, ROUGE,* dan *METEOR.* Penelitian ini menghasilkan sistem terpadu berbasis pembelajaran mendalam yang mampu mendeteksi sekaligus memperbaiki kesopanan teks akademik secara otomatis dalam satu alur pemrosesan yang terintegrasi.

Kata kunci: kesopanan akademik, komunikasi akademik digital, *IndoBERT*, *IndoT5*, transformasi teks

Abstract

The increasing use of digital communication in academic interactions between students and lecturers is often not accompanied by consistent application of language politeness norms, potentially affecting the effectiveness of academic interactions. To date, efforts to enhance language politeness have predominantly relied on manual and subjective evaluation. This study aims to develop an automated system for detecting and improving politeness in Indonesian academic text communication. The proposed approach integrates IndoBERT as a classification model to identify levels of text politeness and IndoT5 as a generative model to transform sentences identified as impolite into more appropriate academic forms. The dataset consists of 6,230 labeled sentences collected through Google Forms, TikTok, and additional synthetic data generated using ChatGPT. Experimental results show that the IndoBERT model achieves an accuracy of 97.11% in classifying academic text politeness, while IndoT5 is capable of transforming impolite sentences into more appropriate academic expressions, as demonstrated by evaluations using BLEU, ROUGE, and METEOR metrics. This study results in an integrated deep learning–based system capable of automatically detecting and improving academic text politeness within a unified processing framework.

Keywords: *academic politeness, digital academic communication, IndoBERT, IndoT5, text transformation*

 ISSN:2302-8149
e-ISSN:2540-9719

1 Pendahuluan

Komunikasi akademik memiliki peranan penting dalam menunjang proses pembelajaran, khususnya dalam interaksi antara mahasiswa dan dosen [1]. Seiring dengan perkembangan teknologi digital, komunikasi akademik tidak lagi terbatas pada pertemuan tatap muka, tetapi juga dilakukan melalui berbagai media digital seperti aplikasi pesan instan dan platform pembelajaran daring [2]. Dalam konteks tersebut, penggunaan bahasa yang santun dan sesuai dengan kaidah akademik menjadi faktor penting agar pesan dapat dipahami dengan baik serta tidak menimbulkan kesalahpahaman.

Namun, dalam praktiknya masih banyak ditemukan komunikasi akademik yang kurang memperhatikan aspek kesopanan berbahasa. Sejumlah penelitian menunjukkan bahwa mahasiswa kerap menggunakan bahasa yang terlalu informal, kalimat yang singkat, serta ungkapan yang tidak sesuai dengan konteks akademik ketika berkomunikasi dengan dosen melalui media digital [3], [4], [5]. Pergeseran pola komunikasi ini berpotensi menurunkan kualitas interaksi akademik dan menciptakan persepsi yang kurang tepat dalam hubungan mahasiswa dan dosen.

Penanganan permasalahan kesopanan berbahasa dalam komunikasi akademik umumnya masih dilakukan secara manual, baik melalui teguran langsung maupun koreksi dari dosen. Pendekatan tersebut memiliki keterbatasan karena bersifat subjektif dan bergantung pada persepsi individu, sehingga konsistensinya sulit dijaga, terutama pada komunikasi digital dengan intensitas yang tinggi [6]. Selain itu, sebagian besar penelitian terdahulu masih berfokus pada kajian deskriptif mengenai kesantunan berbahasa mahasiswa tanpa menyediakan solusi berbasis sistem yang mampu mendeteksi dan memperbaiki teks yang kurang sopan secara otomatis [7].

Seiring dengan berkembangnya pendekatan *Natural Language Processing*, beberapa penelitian telah menerapkan model klasifikasi berbasis transformer seperti IndoBERT untuk memahami konteks bahasa Indonesia dengan baik [8], [9]. Di sisi lain, model generatif seperti *IndoT5* juga telah digunakan untuk melakukan transformasi teks dengan tetap mempertahankan makna kalimat [10]. Meskipun demikian, pendekatan-pendekatan tersebut umumnya masih diterapkan secara terpisah, sehingga belum menyediakan alur terpadu yang mampu melakukan deteksi sekaligus perbaikan kesopanan teks dalam satu sistem otomatis.

Hingga saat ini masih sangat terbatas penelitian yang mengintegrasikan proses deteksi dan perbaikan kesopanan teks dalam satu alur sistem otomatis, khususnya dalam konteks komunikasi akademik Bahasa Indonesia. Sebagian besar penelitian sebelumnya masih berfokus pada klasifikasi atau transformasi teks secara terpisah tanpa menyediakan mekanisme perbaikan bahasa secara langsung. Ketiadaan pendekatan terpadu tersebut menunjukkan adanya celah penelitian yang mendorong perlunya pengembangan sistem otomatis yang mampu melakukan deteksi sekaligus perbaikan kesopanan teks secara terintegrasi.

Berdasarkan permasalahan tersebut, penelitian ini bertujuan untuk mengembangkan sistem otomatis deteksi dan perbaikan kesopanan teks akademik berbasis *IndoBERT* dan *IndoT5*. Sistem yang diusulkan mengintegrasikan model *IndoBERT* untuk mendeteksi tingkat kesopanan teks dan model *IndoT5* untuk memperbaiki kalimat yang terdeteksi tidak sopan agar sesuai dengan kaidah komunikasi akademik. Kontribusi utama penelitian ini terletak pada penyediaan sistem terpadu yang tidak hanya mengidentifikasi tingkat kesopanan teks, tetapi juga secara langsung menghasilkan perbaikan kalimat dalam konteks komunikasi akademik Bahasa Indonesia.

2 Tinjauan Literatur

Kesopanan berbahasa merupakan aspek fundamental dalam komunikasi akademik karena mencerminkan etika, profesionalisme, serta sikap saling menghormati dalam interaksi antara mahasiswa dan dosen. Dalam konteks pendidikan tinggi, kesantunan berbahasa tidak hanya berkaitan dengan pilihan kata, tetapi juga struktur kalimat dan kesesuaian ungkapan dengan situasi komunikasi akademik [3]. Penggunaan bahasa yang santun berperan penting dalam menjaga kualitas interaksi akademik serta menghindari potensi kesalahpahaman dalam penyampaian pesan.

Perkembangan teknologi komunikasi digital turut memengaruhi pola kesantunan berbahasa mahasiswa. Sejumlah penelitian menunjukkan bahwa komunikasi akademik melalui media digital, seperti aplikasi pesan instan dan platform daring, cenderung mendorong penggunaan bahasa yang lebih ringkas dan informal [4], [5]. Dalam praktiknya, mahasiswa masih sering menggunakan kalimat singkat, ungkapan nonformal, serta simbol atau emoji yang kurang sesuai dengan norma komunikasi

 ISSN:2302-8149
e-ISSN:2540-9719

akademik ketika berinteraksi dengan dosen [11]. Kondisi ini menunjukkan adanya pergeseran pola komunikasi yang berpotensi menurunkan tingkat kesopanan dalam konteks akademik.

Sejumlah studi yang mengkaji kesantunan berbahasa mahasiswa di era digital umumnya masih bersifat deskriptif, dengan fokus pada identifikasi bentuk-bentuk ketidaksantunan serta faktor penyebabnya [6], [7]. Meskipun penelitian-penelitian tersebut memberikan gambaran yang komprehensif mengenai fenomena kesantunan berbahasa, pendekatan yang ditawarkan umumnya belum menyediakan solusi praktis berbasis sistem yang dapat membantu mahasiswa memperbaiki kesantunan bahasa secara langsung dalam komunikasi akademik.

Seiring dengan perkembangan bidang *Natural Language Processing* (NLP), berbagai pendekatan berbasis pembelajaran mesin dan *deep learning* telah diterapkan untuk menganalisis teks berbahasa Indonesia. NLP banyak digunakan dalam tugas klasifikasi teks, seperti analisis sentimen dan opini pengguna pada media digital [8]. Beberapa penelitian menerapkan metode *machine learning* untuk mengklasifikasikan sentimen dan opini pada media sosial, namun pendekatan tersebut umumnya hanya berfokus pada proses identifikasi atau klasifikasi teks tanpa melibatkan proses perbaikan atau transformasi teks secara otomatis [12], [13].

Model transformer merupakan pendekatan mutakhir dalam NLP yang mampu menangkap konteks bahasa secara lebih mendalam melalui mekanisme *self-attention*. Salah satu model transformer yang banyak digunakan untuk Bahasa Indonesia adalah IndoBERT. Model ini telah terbukti efektif dalam berbagai tugas klasifikasi teks, termasuk analisis sentimen dan klasifikasi pertanyaan, karena kemampuannya dalam memahami konteks kalimat berbahasa Indonesia secara lebih akurat [9], [14], [15]. Namun demikian, penggunaan IndoBERT umumnya terbatas pada tugas klasifikasi, sehingga belum mencakup proses perbaikan atau transformasi teks.

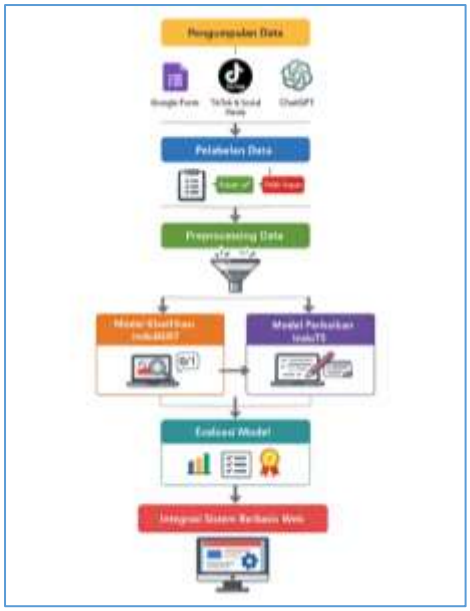
Selain model klasifikasi, model transformer generatif juga berkembang untuk mendukung tugas transformasi teks. *IndoT5* merupakan model *text-to-text* transformer yang dirancang untuk berbagai tugas generatif, termasuk parafrase dan reformulasi kalimat. Penelitian sebelumnya menunjukkan bahwa *IndoT5* mampu menghasilkan teks dengan makna yang setara dengan kalimat masukan, sehingga relevan untuk digunakan dalam perbaikan struktur dan gaya bahasa [10]. Meskipun demikian, penerapan model generatif ini masih jarang dikombinasikan secara langsung dengan model klasifikasi dalam satu sistem terpadu.

Beberapa penelitian telah mencoba menerapkan pendekatan *deep learning* untuk mengubah kalimat tidak sopan menjadi lebih sopan [16]. Namun, penelitian-penelitian tersebut umumnya belum secara khusus mengintegrasikan proses deteksi dan perbaikan kesopanan dalam satu alur sistem yang utuh, terutama dalam konteks komunikasi akademik Bahasa Indonesia. Berdasarkan tinjauan literatur yang telah dilakukan, dapat disimpulkan bahwa masih terdapat celah penelitian dalam pengembangan sistem otomatis yang mampu mendeteksi sekaligus memperbaiki kesopanan teks akademik. Oleh karena itu, penelitian ini mengusulkan sistem terpadu yang mengombinasikan model IndoBERT untuk mendeteksi tingkat kesopanan teks dan model *IndoT5* untuk memperbaiki kalimat yang terdeteksi tidak sopan, sehingga diharapkan dapat meningkatkan kualitas komunikasi akademik mahasiswa secara otomatis.

3 Metode Penelitian

Penelitian ini bertujuan untuk membangun sistem otomatis yang mampu mendeteksi dan memperbaiki tingkat kesopanan teks dalam komunikasi akademik antara mahasiswa dan dosen. Metode penelitian disusun secara sistematis mulai dari tahap pengumpulan data, pelabelan, *preprocessing*, pembangunan model klasifikasi dan perbaikan teks, hingga integrasi sistem berbasis web. Alur penelitian secara keseluruhan ditunjukkan pada Gambar 1.

 ISSN:2302-8149
e-ISSN:2540-9719



Gambar 1 Alur penelitian

3.1 Pengumpulan Data

Data yang digunakan dalam penelitian ini berupa teks komunikasi akademik antara mahasiswa dan dosen dalam Bahasa Indonesia. Data dikumpulkan dari beberapa sumber, yaitu Google Form yang disebarkan kepada mahasiswa, platform TikTok yang memuat unggahan ulang percakapan komunikasi akademik dari media sosial X dan Facebook, serta data tambahan yang dihasilkan menggunakan ChatGPT. Total data yang berhasil dikumpulkan berjumlah 6.230 kalimat, yang terdiri atas 2.220 kalimat dari Google Form, 2.008 kalimat dari TikTok, dan 2.002 kalimat dari ChatGPT.

Data komunikasi akademik yang dikumpulkan mencakup berbagai konteks, antara lain konsultasi skripsi, konfirmasi mengenai menjadi asisten dosen, bimbingan akademik, permohonan izin perkuliahan, serta komunikasi administratif terkait program studi dan konteks akademik lainnya. Proses pengambilan data dilakukan dalam rentang waktu Oktober 2025 hingga Desember 2025.

3.2 Pelabelan Data

Seluruh data yang terkumpul diberi label secara manual ke dalam dua kelas, yaitu label 0 untuk kalimat sopan dan label 1 untuk kalimat tidak sopan. Proses pelabelan dilakukan berdasarkan kaidah kesopanan berbahasa dalam konteks komunikasi akademik, dengan mempertimbangkan pemilihan diksi, struktur kalimat, serta kesesuaian gaya bahasa terhadap relasi mahasiswa–dosen sebagaimana dijelaskan dalam penelitian sebelumnya [3], [5].

Pelabelan manual dipilih untuk memastikan ketepatan interpretasi konteks akademik, mengingat kesopanan bahasa tidak hanya ditentukan oleh bentuk leksikal, tetapi juga oleh konteks sosial dan pragmatik penggunaan bahasa. Proses pelabelan dilakukan oleh satu orang anotator, yaitu peneliti, yang melakukan pelabelan secara manual terhadap seluruh data. Penentuan label sopan dan tidak sopan tidak hanya didasarkan pada bentuk kebahasaan yang baku, tetapi juga mempertimbangkan konteks komunikasi akademik, seperti kejelasan maksud, kesesuaian relasi mahasiswa–dosen, serta kecenderungan mahasiswa menentukan waktu atau agenda bimbingan secara sepihak.

Untuk menjaga kewajaran interpretasi hasil sistem, dilakukan diskusi konseptual dengan seorang ahli psikologi pada tahap evaluasi akhir. Diskusi ini difokuskan pada penilaian kesesuaian luaran sistem terhadap prinsip kesopanan dalam komunikasi akademik, tanpa melakukan perubahan terhadap label data yang telah ditetapkan sebelumnya. Dengan demikian, keterlibatan ahli psikologi berperan sebagai validasi konseptual terhadap keluaran sistem, bukan sebagai bagian dari proses pelabelan data.

3.3 *Preparation* dan *Preprocessing Data*

Tahapan *preparation* dan *preprocessing* dilakukan untuk meningkatkan kualitas data sebelum digunakan dalam proses pelatihan model. Pada tahap *preparation*, data diperiksa untuk memastikan

 ISSN:2302-8149
e-ISSN:2540-9719

kesesuaian antara teks dan label, serta dilakukan penghapusan data yang tidak memiliki teks atau label [17].

Tahapan *preprocessing* meliputi beberapa proses, yaitu *case folding*, penghapusan URL, *mention*, simbol, emoji, serta spasi berlebih [14]. Selain itu, kalimat dengan panjang kurang dari lima karakter dihapus karena dianggap tidak merepresentasikan konteks komunikasi akademik yang memadai. Seluruh proses *preprocessing* dilakukan menggunakan bahasa pemrograman Python dengan dukungan pustaka PyTorch dan Hugging Face Transformers.

3.4 Pembangunan Model Klasifikasi Kesopanan menggunakan *IndoBERT*

Model klasifikasi kesopanan dibangun menggunakan IndoBERT yang telah di-*fine-tuning* pada *dataset* penelitian. *Dataset* hasil *preprocessing* dibagi menjadi tiga bagian, yaitu data latih, data validasi, dan data uji dengan rasio 70% : 15% : 15%, menggunakan metode *stratified split* untuk menjaga keseimbangan distribusi kelas [15].

Proses pelatihan dilakukan selama 3 epoch dengan *batch size* 8 untuk data latih dan 16 untuk data validasi serta pengujian. Evaluasi dilakukan pada setiap epoch, dan model terbaik dipilih berdasarkan nilai akurasi tertinggi. Pemanfaatan IndoBERT dipilih karena terbukti efektif dalam berbagai tugas klasifikasi teks Bahasa Indonesia [9], [14], [15].

Penelitian ini tidak mengeksplorasi variasi *optimizer* dan *learning rate*, karena fokus utama terletak pada integrasi sistem deteksi dan perbaikan kesopanan teks. Proses pelatihan dilakukan selama tiga epoch berdasarkan eksperimen awal yang menunjukkan penurunan performa pada data validasi ketika jumlah epoch ditingkatkan. Oleh karena itu, mekanisme *early stopping* tidak diterapkan.

3.5 Pembangunan Model Perbaikan Teks menggunakan *IndoT5*

Model perbaikan teks dibangun menggunakan *IndoT5* dengan pendekatan *sequence-to-sequence* untuk mentransformasi kalimat yang terdeteksi tidak sopan menjadi lebih sopan tanpa mengubah makna awal. *Dataset* dibagi menjadi data latih, data validasi, dan data uji dengan rasio 80% : 10% : 10% menggunakan metode *stratified split* [14].

Proses pelatihan dilakukan selama 3 epoch dengan *batch size* 4 untuk data latih dan 8 untuk data validasi. Evaluasi dilakukan untuk mengukur kemampuan model dalam mempertahankan makna dan struktur kalimat hasil perbaikan, sebagaimana direkomendasikan pada penelitian sebelumnya [10].

3.6 Evaluasi Model

Evaluasi dilakukan terhadap dua model yang dibangun, yaitu model klasifikasi dan model perbaikan teks. Evaluasi model klasifikasi dilakukan menggunakan *confusion matrix*, *precision*, *recall*, dan *F1-score* [9], [18]. Sementara itu, evaluasi model perbaikan teks dilakukan menggunakan metrik *BLEU*, *ROUGE-1*, *ROUGE-2*, *ROUGE-L*, dan *METEOR* untuk mengukur kesesuaian semantik dan kualitas transformasi kalimat [10].

3.7 Integrasi Sistem Berbasis Web

Tahapan akhir penelitian adalah integrasi kedua model ke dalam sebuah sistem berbasis web. Sistem ini dibangun menggunakan bahasa pemrograman Python dengan framework FastAPI untuk *backend*, sedangkan *frontend* menggunakan HTML, CSS (Bootstrap), dan JavaScript. *Backend* memuat langsung model *IndoBERT* dan *IndoT5*, sehingga pemrosesan klasifikasi dan perbaikan teks dilakukan secara lokal di server. Alur kerja sistem secara rinci adalah sebagai berikut:

a. Pengguna menginput teks komunikasi akademik melalui antarmuka *frontend*.
b. *Frontend* mengirim data teks ke *backend* menggunakan format JSON.
c. *Backend* menerima *request* melalui FastAPI dan memvalidasi apakah teks berkaitan dengan konteks akademik. Jika teks tidak relevan, sistem akan menampilkan pesan bahwa teks tersebut di luar konteks akademik.
d. Jika teks relevan, *backend* melakukan *preprocessing*, kemudian model *IndoBERT* mengklasifikasikan tingkat kesopanan teks.
e. Jika teks sudah sopan, sistem langsung menampilkan hasil ke pengguna.
f. Jika teks diklasifikasi menjadi tidak sopan, *backend* memproses teks tersebut menggunakan IndoT5 untuk memperbaiki kalimat menjadi lebih sopan tanpa mengubah makna aslinya.

 ISSN:2302-8149
e-ISSN:2540-9719

g. *Backend* mengirim hasil perbaikan kembali ke *frontend* untuk ditampilkan sebagai *output* akhir kepada pengguna.

Integrasi ini memungkinkan sistem bekerja secara *end-to-end*, mulai dari identifikasi kalimat tidak sopan hingga menghasilkan kalimat yang lebih sesuai dengan kaidah komunikasi akademik. Dengan desain ini, sistem tidak hanya mendukung proses komunikasi akademik yang sopan, tetapi juga memudahkan mahasiswa dalam memahami dan mempraktikkan penggunaan bahasa yang sesuai konteks.

4 Hasil dan Pembahasan

Bab ini menyajikan hasil pengujian dan pembahasan sistem deteksi serta perbaikan kesopanan teks akademik berbasis *IndoBERT* dan *IndoT5*. Penyajian hasil disusun mengikuti tahapan metodologi penelitian, mulai dari karakteristik *dataset*, hasil *preprocessing*, evaluasi model klasifikasi kesopanan, evaluasi model perbaikan teks, hingga integrasi sistem. Penyusunan ini dimaksudkan untuk menunjukkan keterkaitan yang jelas antara metode yang digunakan dan hasil yang diperoleh.

Dataset yang digunakan dalam penelitian ini terdiri atas 6.230 data komunikasi akademik mahasiswa–dosen dalam Bahasa Indonesia yang telah melalui proses pelabelan dan pembersihan data. Distribusi *dataset* menunjukkan jumlah data yang seimbang antara kelas kalimat sopan dan tidak sopan, sebagaimana disajikan pada Tabel 1. Keseimbangan ini menjadi penting untuk memastikan evaluasi model klasifikasi dilakukan secara adil tanpa bias terhadap salah satu kelas.

Tabel 1 Distribusi *dataset* berdasarkan label kesopanan

Label	Kategori	Jumlah Data
1	Tidak sopan	3115
0	Sopan	3115
Total		**6230**

Tabel 2 menyajikan contoh pasangan data yang digunakan dalam pelatihan model perbaikan teks. Kolom *Bentuk Teks Sopan atau Tidak Sopannya* menampilkan pasangan kalimat dari *dataset* yang merepresentasikan hubungan antara bentuk kalimat tidak sopan dan bentuk kalimat sopannya. Untuk data berlabel tidak sopan (label 1), pasangan yang ditampilkan adalah bentuk kalimat sopannya. Sebaliknya, untuk data berlabel sopan (label 0), pasangan yang ditampilkan merupakan contoh bentuk kalimat tidak sopan yang terdapat dalam *dataset* sebagai pasangan pelatihan. Pasangan ini digunakan untuk membantu model mempelajari pola transformasi kesopanan bahasa dan bukan merupakan hasil keluaran sistem.

Tabel 2 Contoh pasangan teks sopan dan tidak sopan dalam *dataset* pelatihan

Teks	Label	Bentuk Teks Sopan atau Tidak Sopannya
Bu, saya udah ngirim tugasnya, cepet diperiksa ya	1	Bu, saya sudah mengirimkan tugasnya. Mohon kesediaannya untuk diperiksa jika berkenan. Terima kasih sebelumnya.
pak saya telat soalnya macet	1	Mohon maaf, Pak. Saya datang agak terlambat karena kondisi jalan sedang macet.
bu kapan nilai saya keluar? lama banget	1	Permisi, Bu. Saya ingin menanyakan apakah nilai mata kuliah ini sudah keluar? Terima kasih sebelumnya.
Mohon informasi dari Ibu terkait jadwal pengumpulan revisi terakhir.	0	bu deadline revisi terakhir kapan ya?
Mohon arahan Ibu mengenai tata cara pengajuan sidang skripsi.	0	Bu, gimana sih cara daftar sidang skripsi? 😅
Mohon informasi tambahan dari Ibu terkait poin penilaian yang belum terpenuhi.	0	Bu, yang kurang di nilai saya tuh bagian mana ya?

 ISSN:2302-8149
e-ISSN:2540-9719

Hasil *preprocessing* data menunjukkan adanya pengurangan jumlah data dari 6.230 menjadi 6.221 kalimat, sebagaimana ditampilkan pada Tabel 3. Pengurangan ini disebabkan oleh penghapusan data kosong, kalimat yang hanya berisi simbol atau emoji, serta kalimat dengan panjang kurang dari lima karakter yang dianggap tidak merepresentasikan konteks komunikasi akademik secara memadai. Contoh hasil *preprocessing* pada Tabel 4 memperlihatkan bahwa proses pembersihan data berhasil menghilangkan elemen non-linguistik tanpa mengubah makna utama kalimat. Data hasil *preprocessing* ini kemudian digunakan sebagai input untuk pelatihan model klasifikasi kesopanan berbasis IndoBERT, sejalan dengan praktik umum dalam pengolahan teks Bahasa Indonesia [14], [17].

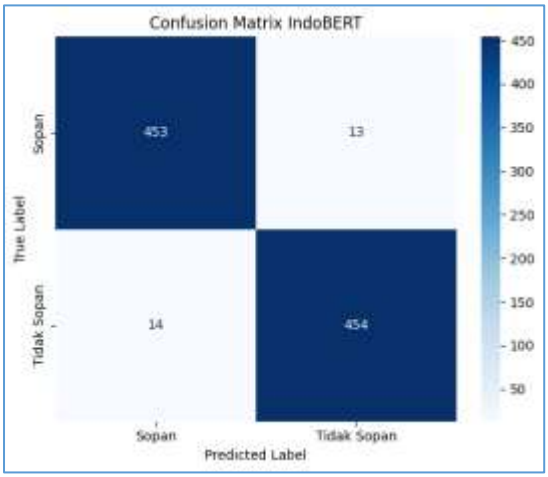
Tabel 3 Perbandingan *dataset* sebelum dan sesudah *preprocessing*

Tahapan	Total Data	Sopan	Tidak Sopan
Data Awal	6230	3115	3115
Setelah *preparation* dan *preprocessing*	6221	3107	3114

Tabel 4 Contoh teks sebelum dan sesudah *preprocessing*

Kalimat Asli	Kalimat Setelah *Preprocessing*
Bu, gimana sih cara daftar sidang skripsi? 😅	bu gimana sih cara daftar sidang skripsi
Bu, saya udah ngirim tugasnya, cepet diperiksa ya	bu saya udah ngirim tugasnya cepet diperiksa ya

Evaluasi performa model klasifikasi kesopanan menggunakan *IndoBERT* dilakukan pada data uji sebanyak 934 data. Hasil *confusion matrix* yang ditampilkan pada Gambar 2 menunjukkan bahwa model mampu mengklasifikasikan sebagian besar data uji dengan benar, baik pada kelas kalimat sopan maupun tidak sopan. Dari total data uji, sebanyak 454 kalimat tidak sopan dan 453 kalimat sopan berhasil diprediksi secara tepat oleh model. Hasil evaluasi kuantitatif pada Tabel 5 menunjukkan bahwa model *IndoBERT* mencapai nilai *accuracy* sebesar 97,11%, dengan nilai *precision*, *recall*, dan *F1-score* yang sama besar.



Gambar 2 *Confusion matrix* evaluasi model klasifikasi kesopanan berbasis *IndoBERT*

Tabel 5 Performa model *IndoBERT* pada tugas klasifikasi kesopanan

Metrik	Nilai
Accuracy	97,11%
Precision	97,11%
Recall	97,11%
F1-Score	97.11%

 ISSN:2302-8149
e-ISSN:2540-9719

Nilai performa yang konsisten pada seluruh metrik tersebut menunjukkan bahwa model mampu memahami konteks kesopanan bahasa Indonesia dalam komunikasi akademik dengan baik. Kesetaraan nilai *precision*, *recall*, dan *F1-score* mengindikasikan bahwa model tidak bias terhadap salah satu kelas, yang didukung oleh distribusi data uji yang seimbang. Jika dibandingkan dengan penelitian sebelumnya yang menggunakan IndoBERT untuk berbagai tugas klasifikasi teks Bahasa Indonesia dengan tingkat akurasi pada rentang 82% hingga 97% [9], [14], hasil penelitian ini menunjukkan performa yang kompetitif. Pencapaian ini dipengaruhi oleh penggunaan *dataset* yang spesifik pada konteks komunikasi akademik, proses pelabelan manual yang mempertimbangkan aspek pragmatik bahasa, serta penerapan *stratified split* untuk menjaga keseimbangan kelas selama proses pelatihan.

Setelah proses deteksi kesopanan, kalimat yang terklasifikasi sebagai tidak sopan diproses lebih lanjut menggunakan model *IndoT5* untuk dilakukan perbaikan teks. Model ini dirancang untuk mentransformasi gaya bahasa kalimat menjadi lebih sopan tanpa mengubah makna awal pesan. Contoh hasil perbaikan teks menggunakan IndoT5 ditampilkan pada Tabel 6. Berdasarkan contoh tersebut, terlihat bahwa model mampu menambahkan unsur kesantunan, seperti pembukaan komunikasi, pemilihan diksi yang lebih formal, serta struktur kalimat yang lebih sesuai dengan konteks akademik.

Tabel 6 Contoh hasil perbaikan teks menggunakan model *IndoT5*

Kalimat Awal	Hasil Perbaikan
Assalamualaikum, Bu. Saya mau bimbingan. Bisa kapan ya?	Assalamualaikum, Bu. Saya ingin melakukan bimbingan. Apakah Ibu memiliki waktu luang?
Bu, tugas ini dinilainya kapan ya? Saya belum lihat nilainya di Google Classroom.	Assalamualaikum Bu, saya ingin menanyakan hasil penilaian tugas ini di Google Classroom.
Pak, UTS nya kok jadwalnya beda ya sama yang di formulir?	Assalamualaikum, Pak. Mohon izin bertanya, apakah ada perbedaan jadwal untuk UTS dengan yang ada di formulir?
Misi pak saya mau bimbingan skripsi	Assalamualaikum, Pak. Mohon izin, saya ingin melakukan bimbingan skripsi. Apakah Bapak berkenan?

Evaluasi model *IndoT5* dilakukan menggunakan metrik *BLEU*, *ROUGE-1*, *ROUGE-2*, *ROUGE-L*, dan *METEOR*, dengan hasil sebagaimana disajikan pada Tabel 7. Nilai *BLEU* yang relatif rendah menunjukkan bahwa kesamaan n-gram antara teks hasil perbaikan dan teks referensi tidak selalu tinggi, yang wajar mengingat perbaikan kesopanan lebih menekankan pada perubahan gaya bahasa daripada kesamaan struktur literal. Sebaliknya, nilai *ROUGE* dan *METEOR* yang cukup tinggi mengindikasikan bahwa model mampu mempertahankan kesesuaian semantik antara kalimat awal dan kalimat hasil perbaikan. Temuan ini sejalan dengan penelitian sebelumnya yang menyatakan bahwa metrik *ROUGE* dan *METEOR* lebih representatif dalam mengevaluasi tugas parafrase dan transformasi gaya bahasa dibandingkan *BLEU* [10].

Tabel 7 Hasil Evaluasi Model *IndoT5*

Metrik	Nilai
BLEU	0,1852
ROUGE-1	0,5816
ROUGE-2	0,3325
ROUGE-L	0,5333
METEOR	0,4601

Hasil penelitian ini memperluas pemanfaatan *IndoT5* yang sebelumnya banyak digunakan untuk tugas parafrase Bahasa Indonesia [10], dengan mengintegrasikannya ke dalam sistem yang secara otomatis mendeteksi dan memperbaiki kesopanan teks. Integrasi antara model deteksi kesopanan dan

 ISSN:2302-8149
e-ISSN:2540-9719

model perbaikan teks memungkinkan sistem bekerja secara *end-to-end*, mulai dari identifikasi kalimat tidak sopan hingga menghasilkan kalimat yang lebih sesuai dengan kaidah komunikasi akademik.

Tahap akhir penelitian adalah integrasi kedua model ke dalam sebuah sistem berbasis web. Sistem ini menerima input berupa teks komunikasi akademik dari pengguna, kemudian melakukan deteksi tingkat kesopanan secara otomatis. Apabila kalimat terdeteksi sopan, sistem langsung menampilkan hasil tanpa perubahan. Sebaliknya, jika kalimat terdeteksi tidak sopan, sistem akan menghasilkan versi perbaikan menggunakan model *IndoT5* sebelum ditampilkan kepada pengguna. Tampilan antarmuka sistem ditunjukkan pada Gambar 3. Integrasi ini membedakan penelitian ini dari studi sebelumnya yang umumnya memisahkan proses klasifikasi dan transformasi teks [8], [16], serta menunjukkan kontribusi sistem secara praktis dalam membantu mahasiswa menyusun komunikasi akademik yang lebih santun dan sesuai konteks.

Gambar 3 Antarmuka sistem deteksi dan perbaikan kesopanan teks akademik berbasis web

5 Kesimpulan

Penelitian ini telah mencapai tujuannya, yaitu mengembangkan sistem otomatis yang mampu mendeteksi dan memperbaiki kesopanan teks dalam komunikasi akademik berbahasa Indonesia secara terintegrasi. Model *IndoBERT* menunjukkan performa klasifikasi yang sangat tinggi dengan *accuracy*, *precision*, *recall*, dan *F1-score* masing-masing sebesar 97,11%, sedangkan model *IndoT5* memperoleh skor *BLEU* 0,1852, *ROUGE-1* 0,5816, *ROUGE-2* 0,3325, *ROUGE-L* 0,5333, dan *METEOR* 0,4601, yang menunjukkan kemampuannya dalam mentransformasikan kalimat tidak sopan menjadi lebih santun tanpa mengubah makna utama. Secara teknis, penelitian ini membuktikan bahwa integrasi model klasifikasi dan generatif berbasis pembelajaran mendalam dapat diimplementasikan secara efektif dalam satu sistem *end-to-end* untuk pemrosesan teks Bahasa Indonesia. Secara praktis, sistem ini berpotensi menjadi alat bantu sekaligus media edukasi bagi mahasiswa dalam meningkatkan kualitas dan kesopanan komunikasi akademik dengan dosen. Penelitian selanjutnya dapat diarahkan pada pelaksanaan *usability testing* secara langsung kepada pengguna, pengembangan strategi *prompt engineering* untuk meningkatkan kualitas generasi teks, serta pengujian lintas institusi guna memperkuat generalisasi dan skalabilitas sistem.

Referensi

[1] N. P. Lestari, "Pengaruh Pola Komunikasi Mahasiswa dengan Dosen Pembimbing Akademik dan Motivasi Belajar Mahasiswa Pendidikan Ilmu Pengetahuan Sosial di UIN Malang," *Dinamika Sosial: Jurnal Pendidikan Ilmu Pengetahuan Sosial*, vol. 1, no. 1, pp. 1–11, 2022, doi: 10.18860/dsjpips.v1i1.1009.

[2] K. Abidin and Wandi, "Etika Komunikasi antara Mahasiswa dan Dosen dalam Interaksi Akademik melalui Media Digital," *MEDIALOG: Jurnal Ilmu Komunikasi*, vol. VI, no. I, 2023.

[3] N. H. Mahas, Muhsyanur, and S. Verlin, "Pola Kesantunan Berbahasa Indonesia dalam Komunikasi Akademik Mahasiswa Universitas Islam As'adiyah Sengkang: Analisis Sosiopragmatik," *Aksentuasi: Jurnal Ilmiah Pendidikan Bahasa dan Sastra Indonesia*, vol. 6, no. 1, pp. 138–149, 2025.

 ISSN:2302-8149
e-ISSN:2540-9719

[4] H. I. Harahap, N. Widiawati, S. Putri, P. D. Siregar, and E. Syahirah, "Etika Mahasiswa terhadap Dosen secara Online Maupun Offline," 2023.
[5] E. R. Eksan, A. Hafid, and T. Y. Putra, "Kesantunan Berbahasa Mahasiswa terhadap Dosen di UNIMUDA Sorong: Tinjauan Pragmatik," *FRASA: Jurnal Keilmuan, Bahasa, Sastra, dan Pengajarannya*, vol. 2, no. 1, pp. 16–23, 2021.
[6] P. Astiti and D. M. Raharja, "Peran Komunikasi dalam Pendidikan Era Digital," *Gandiwa: Jurnal Komunikasi*, vol. 03, no. 02, pp. 41–48, 2023, doi: 10.30998/g.v3i2.2514.
[7] R. Ningsih and F. Fatmawati, "Realitas Kesantunan Berbahasa Gen-Z di Era Digital," *Jurnal Onoma: Pendidikan Bahasa, dan Sastra*, vol. 10, no. 1, pp. 215–224, 2024.
[8] R. A. Fikri and R. Muliono, "Sentiment Towards Social Media Politeness Ambassadors: A Case Study Using the Naive Bayes Method," *JOURNAL OF INFORMATICS AND TELECOMMUNICATION ENGINEERING*, vol. 8, no. 3Spc, pp. 127–135, 2025, doi: 10.31289/jite.v8i3Spc.14404.
[9] F. Baharuddin and M. F. Naufal, "Fine-tuning IndoBERT for Indonesian exam question classification based on Bloom's Taxonomy," *Journal of Information Systems Engineering and Business Intelligence*, vol. 9, no. 2, pp. 253–263, Oct. 2023, doi: 10.20473/jisebi.9.2.253-263.
[10] W. Puspitasari, D. Ramdani, and A. M. Maulana, "IndoT5 (Text-to-Text Transfer Transformer) Algorithm for Paraphrasing Indonesian Language Islamic Sermon Manuscripts," *Khazanah Journal of Religion and Technology*, vol. 2, no. 2, pp. 63–73, Jan. 2024, doi: 10.15575/kjrt.v2i2.1093.
[11] S. Z. Kurniadini and T. D. Putranto, "Pola Komunikasi Interpersonal Mahasiswa Indonesia (Review of Literature)," *Jurnal Ilmu Komunikasi*, vol. 15, no. 2, p. 84, 2025, doi: 10.36989/didaktik.v10i3.4130.
[12] M. N. Akhfasy, A. K. Ningsih, and R. Ilyas, "Analisis Sentimen Pengguna Instagram terhadap Komentar Karen's Diner Menggunakan Metode Algoritma Support Vector Machine dengan Operator Seleksi Fitur Information Gain," *Jurnal Mahasiswa Teknik Informatika*, vol. 8, no. 3, pp. 2699–2705, 2024, doi: 10.36040/jati.v8i3.9533.
[13] A. K. Sari, A. Irsyad, D. N. Aini, Islamiyah, and S. E. Ginting, "Analisis Sentimen Twitter Menggunakan Machine Learning untuk Identifikasi Konten Negatif," *Adopsi Teknologi dan Sistem Informasi (ATASI)*, vol. 3, no. 1, pp. 64–73, Jun. 2024, doi: 10.30872/atasi.v3i1.1373.
[14] M. F. Kono, I. N. Fajri, and Y. Pristyanto, "Public Sentiment Analysis on Corruption Issues in Indonesia Using IndoBERT Fine-Tuning, Logistic Regression, and Linear SVM," *Journal of Applied Informatics and Computing (JAIC)*, vol. 9, no. 5, pp. 2616–2628, 2025, doi: doi.org/10.30871/jaic.v9i5.10537.
[15] Tarwoto, R. Nugroho, N. Azka, and W. S. R. Graha, "Analisis Sentimen Ulasan Aplikasi Mobile JKN di Google PlayStore Menggunakan IndoBERT," vol. 3, no. 2, pp. 495–505, doi: 10.35870/jtik.v9i2.3340.
[16] A. Christiant and A. R. Pratama, "Implementasi Deep Learning untuk Mengubah Kalimat Tidak Sopan menjadi Sopan," *AUTOMATA*, vol. 3, no. 1, 2022.
[17] D. Irmayanti and T. I. Hermanto, "Development of Hybrid K-Means DBSCAN Algorithm for Optimization of Landslide-Prone Area Clusters based on Web-GIS," *Sistemasi: Jurnal Sistem Informasi*, vol. 15, no. 1, pp. 109–120, 2026, doi: 10.32520/stmsi.v15i1.5671.
[18] D. Nuryadi *et al.*, "Fine-Tuning IndoBERT untuk Analisis Sentimen Ulasan Pengguna Aplikasi Tiket.com di Google Play Store," *Jurnal Mahasiswa Teknik Informatika*, vol. 9, no. 2, pp. 3577–3583, 2025, doi: 10.36040/jati.v9i2.13204.