P- ISSN: 2355-6684, E-ISSN: 2776-6640
https://ojs.unpkediri.ac.id/index.php/noe

Klasifikasi Artikel Ilmiah Terindeks Garuda dengan metode cGAN dan BERT

Ainun Dea Rahayu[1*], Badie'ah[2]
Universitas Islam Sultan Agung Semarang[1,2]
ainundea@std.unissula.ac.id[1], badie'ah.assegaf@unissula.ac.id[2]
***Corresponding author:* Ainun Dea Rahayu**

Abstrak

Jumlah publikasi karya ilmiah terus meningkat, mencapai sekitar 2,6 hingga 3 juta artikel setiap tahun. Peningkatan ini menimbulkan tantangan besar dalam otomatisasi proses pengelompokan dan klasifikasi bidang ilmu, terutama karena distribusi data antar kelas yang tidak merata membuat proses klasifikasi menjadi semakin kompleks. Untuk mengatasi hal ini, penelitian ini menggabungkan dua pendekatan yaitu Conditional Generative Adversarial Network (cGAN) dan Bidirectional Encoder Representations from Transformers (BERT). Model cGAN dimanfaatkan untuk menciptakan data sintetis bagi kelas minoritas dengan pendekatan TF-IDF, sementara BERT digunakan untuk klasifikasi berbasis pemahaman konteks. Data hasil sintesis kemudian digabung dengan data asli dan diuji menggunakan metrik seperti akurasi, presisi, recall, dan F1-score. Hasilnya, metode ini mampu mencapai akurasi hingga 87% serta memperbaiki keseimbangan distribusi prediksi antar bidang ilmu. Pendekatan cGAN-BERT ini diharapkan menjadi solusi untuk mengatasi masalah data tidak seimbang dan memiliki potensi besar untuk diterapkan pada sistem klasifikasi otomatis artikel ilmiah.

Kata Kunci : *Augmentasi Data, cGAN, BERT, Klasifikasi bidang ilmu, Artikel ilmiah*

A. PENDAHULUAN

Dalam era digitalisasi, jumlah publikasi karya ilmiah terus meningkat secara signifikan, mencapai 2,6 hingga 3 juta artikel per tahun. Lonjakan ini memberi peluang besar bagi perkembangan ilmu pengetahuan, tetapi juga menciptakan tantangan besar bagi peneliti untuk menemukan publikasi yang relevan dengan bidang kajian mereka. Kecepatan dan ketepatan dalam memperoleh data terkini sangat dipengaruhi oleh kemajuan teknologi (Dameani, 2021). Salah satu teknologi yang berkembang pesat adalah kecerdasan buatan (*Artificial Intelligence*/AI), yang telah memberikan dampak signifikan di berbagai bidang, termasuk pengembangan sistem klasifikasi (Hajkowicz dkk., 2023). Dalam dunia akademik, sistem klasifikasi berperan penting untuk mempermudah peneliti menemukan publikasi ilmiah yang sesuai dengan topik penelitian mereka (Li dkk., 2024).

Di Indonesia, salah satu platform utama untuk mengakses publikasi akademik adalah Garba Rujukan Digital (Garuda). Platform ini mengindeks ribuan jurnal dari berbagai disiplin ilmu, baik nasional maupun internasional, dan memuat artikel dari peneliti dengan beragam tingkat akreditasi (Sa'adah, 2022). Namun, kemampuan sistem informasi Garuda untuk mengklasifikasikan artikel secara otomatis masih sangat terbatas. Padahal, klasifikasi artikel yang akurat penting untuk memudahkan pencarian, penyusunan metadata, analisis tren keilmuan, hingga pemetaan riset nasional.

Permasalahan ini semakin kompleks karena distribusi publikasi antar bidang penelitian tidak seimbang, di mana beberapa kategori jauh lebih dominan dibandingkan yang lain. Ketidakseimbangan ini berdampak signifikan pada performa model klasifikasi, terutama dalam konteks *deep learning*, yang cenderung bias terhadap kelas mayoritas (Hafiz dan Sudarmilah, 2023). Akibatnya, topik-topik minoritas sering kali kurang terwakili dalam hasil pencarian maupun rekomendasi.

Untuk mengatasi masalah tersebut, dibutuhkan pendekatan modern berbasis *Natural Language* Processing (NLP) dan *Deep Learning* (DL) yang tidak hanya memahami konteks bahasa alami tetapi juga mampu menyeimbangkan distribusi data pelatihan. Salah satu metode yang potensial adalah *Conditional Generative Adversarial Network* (cGAN), yang mampu menghasilkan sampel sintetis dari kelas minoritas melalui interaksi kompetitif antara *generator* dan *discriminator* (Bhat dan Nanjundegowda, 2025). Metode ini efektif untuk mengatasi ketidakseimbangan data dan meningkatkan kinerja model, khususnya pada kelas minoritas (Wang dkk., 2021).

Di sisi lain, *Bidirectional Encoder Representations from Transformers* (BERT) telah terbukti unggul dalam memahami konteks bahasa alami secara dua arah, baik dari kiri ke kanan maupun sebaliknya (Rogers dkk., 2020). Dalam konteks klasifikasi artikel ilmiah, BERT dapat menganalisis judul dan abstrak untuk menentukan kategori penelitian secara lebih akurat (Adhikari dkk., 2020). Namun, performa BERT tetap bergantung pada distribusi data pelatihan yang seimbang.

Sayangnya, riset yang mengintegrasikan BERT dengan teknik *data balancing* berbasis cGAN pada data teks berbahasa Indonesia masih jarang dilakukan, terutama dalam konteks klasifikasi artikel ilmiah lokal. Oleh karena itu, penelitian ini menjadi langkah awal penting untuk mengeksplorasi potensi integrasi BERT dan cGAN pada sumber daya digital nasional seperti Garuda.

Berdasarkan latar belakang dan urgensi tersebut, penelitian ini bertujuan untuk mengembangkan *pipeline* klasifikasi dokumen ilmiah berbasis BERT dan cGAN pada data asli dari Garuda. Dengan pendekatan ini, diharapkan sistem klasifikasi dapat meningkatkan akurasi prediksi, khususnya pada kelas minoritas, sekaligus mendorong pemanfaatan AI mutakhir untuk mendukung pengelolaan pengetahuan dan riset nasional secara lebih efektif.

.

B. LANDASAN TEORI

1. Sistem klasifikasi

Sistem klasifikasi adalah mekanisme pengelompokan entitas berdasarkan kesamaan untuk memudahkan pengelolaan dan pencarian informasi secara sistematis dan terstruktur (Anggraeni dkk., 2021). Fungsi utama sistem klasifikasi dalam konteks ilmiah meliputi mempermudah pencarian informasi, mengorganisir data dalam jumlah besar, meningkatkan efisiensi penemuan kembali informasi, serta memberikan struktur semantik yang jelas antar dokumen (Syarifudin 2022). Integrasi kecerdasan buatan, seperti Natural Language Processing (NLP), GAN, dan BERT, semakin memperkuat sistem klasifikasi digital yang adaptif dan otomatis (Shah dkk., 2023).

2. Generative Adversarial Network (GAN)

GAN adalah model deep learning yang terdiri dari dua jaringan, yaitu generator yang menghasilkan data sintetis dan discriminator yang membedakan data asli dan palsu. GAN efektif untuk augmentasi data, terutama pada dataset yang tidak seimbang, karena dapat menghasilkan data dengan kemiripan semantik tinggi tanpa memerlukan label manual (Suprapti dkk., 2023).

3. Conditional GAN (cGAN)

cGAN merupakan pengembangan GAN yang menggabungkan informasi kondisi (label) ke dalam proses generasi data, memungkinkan kontrol yang lebih spesifik dan peningkatan kualitas data sintetik yang dihasilkan, terutama untuk augmentasi data terbatas (Ma dan Qu 2022).

4. Transformers

Transformers adalah arsitektur jaringan saraf yang menggunakan mekanisme self-attention untuk memahami konteks jangka panjang dalam data sekuensial. Model ini terdiri dari encoder dan decoder yang efisien dalam berbagai tugas NLP seperti klasifikasi teks dan text generation (Islam dkk. 2023).

5. Bidirectional Encoder Representations from Transformers (BERT)

BERT adalah model pretrained berbasis transformer yang membaca konteks dua arah sekaligus sehingga mampu memahami makna kata secara kontekstual. BERT melalui tahap pre-training dan fine-tuning, digunakan luas dalam tugas klasifikasi teks dan pemahaman bahasa (Nayla dkk., 2023).

6. IndoBERT

IndoBERT adalah model BERT yang diadaptasi khusus untuk bahasa Indonesia, dilatih dengan dataset besar berbahasa Indonesia, dan digunakan untuk tugas NLP di Indonesia tanpa perlu pelatihan ulang dari nol (Widiansyah dkk.,2021).

7. Garba Rujukan Digital (Garuda)

Garuda adalah portal referensi digital nasional yang menyediakan akses terhadap dokumen akademik seperti skripsi, tesis, dan artikel ilmiah, yang mendukung penyebaran karya ilmiah dan pengelolaan informasi di Indonesia (Wijaya dan Negara 2022).

C. METODE PENELITIAN

Penelitian ini menggunakan metode klasifikasi berbasis *machine learning Natural Language* Processing (*NLP*) yaitu *Conditional Generative Adversarial* Network (cGAN) dan *Bidirectional Encoder Representations from Transformers* (BERT) digabungkan untuk menghasilkan model klasifikasi yang lebih akurat terhadap artikel ilmiah yang terindeks Garuda.

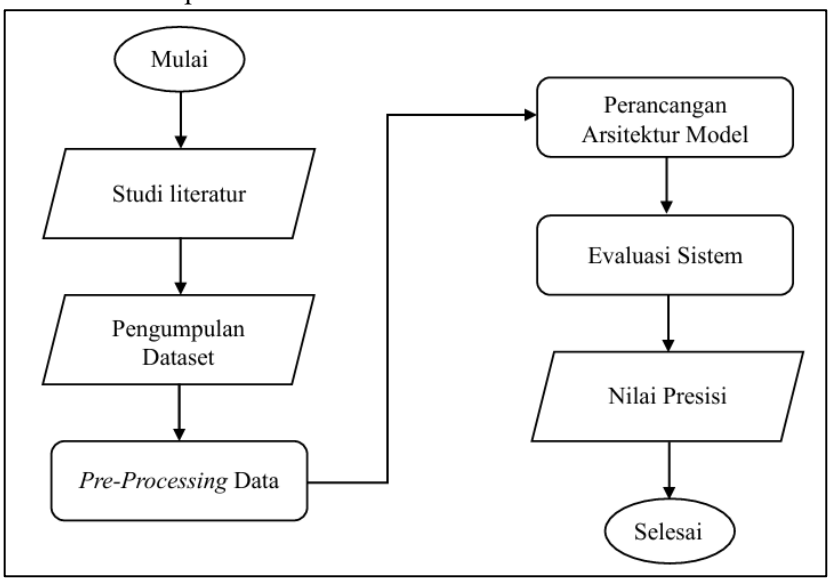
cGAN digunakan untuk menangani ketidakseimbangan data dalam dataset dengan menghasilkan data sintetik yang menyerupai data asli dari kategori minoritas, sehingga memperbesar dan memperbaiki distribusi dataset. BERT digunakan untuk melatih model klasifikasi, di mana representasi teks yang kaya dari BERT akan membantu memahami konteks artikel dengan lebih baik.

P- ISSN: 2355-6684, E-ISSN: 2776-6640
https://ojs.unpkediri.ac.id/index.php/noe

Integrasi keduanya dilakukan dengan memanfaatkan data hasil augmentasi dari cGAN sebagai input tambahan dalam pelatihan BERT, sehingga data yang lebih beragam dapat meningkatkan kemampuan generalisasi dan akurasi klasifikasi model secara keseluruhan.

Penggabungan ini memberikan keuntungan utama, yaitu memanfaatkan data tak berlabel dan data sintetik untuk meningkatkan performa BERT dalam klasifikasi dokumen, bahkan dengan dataset berlabel yang terbatas (Croce dkk., 2020).

Berikut merupakan gambaran alur *flowchart* untuk proses penelitian ini dari pengumpulan data hingga diperoleh hasil untuk implementasi metode cGAN - BERT:



Gambar 1. Alur Penelitian

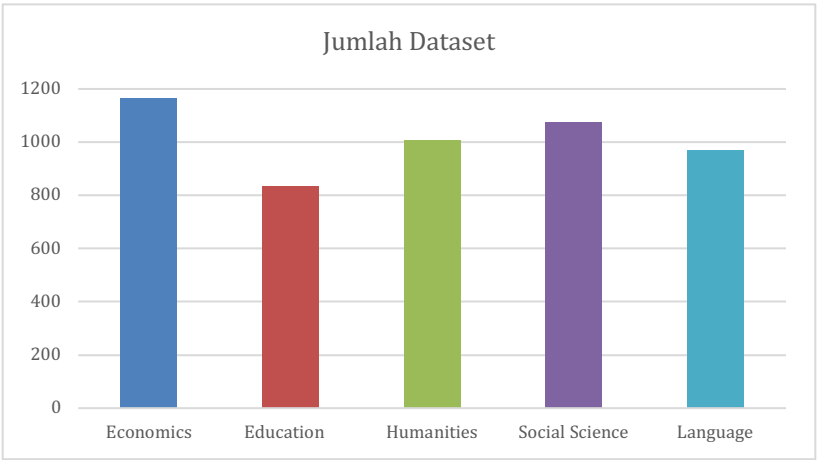
Gambar 1 menunjukan alur kerja penelitian sistem klasifikasi artikel ilmiah terindeks Garuda dengan metode cGAN dan BERT, proses ini meliputi:

1. Studi Literatur

Dengan melakukan studi literasi, penulis dapat mempelajari teori *text mining* tentang Preprocessing *Text* baik berupa proses pada *Natural Language Processing* (NLP) serta implementasi metode cGAN dan BERT dari berbagai sumber literasi seperti tesis, jurnal, makalah, skripsi, ataupun situs-situs *website* yang akan diulas untuk mempelajari teori tersebut.

2. Pengumpulan Dataset

Teknik pengumpulan data pada penelitian ini dilakukan dengan metode *web scrapping*, yaitu teknik ekstrasi data dari halaman *web* secara otomatis menggunakan bantuan bahasa pemrograman. Proses ini mengidentifikasi struktur web terutama pada bagian yang menampilkan daftar artikel, judul, dan abstrak dari *platform* Garuda yang kemudian akan diambil sebagai bahan dataset penelitian.



Gambar 2. *chart database* hasil *scraping*

Gambar 2 merupakan *chart* dataset setelah *scraping* data. Dengan mengambil judul dan abstrak publikasi dari artikel ilmiah total data hasil *scraping* adalah 4766 artikel dari 20 jurnal terindeks garuda dengan jumlah per-datanya yaitu *Economics* 1166 data, *Education* 752 data, *Humanities* 1008 data, *Social Science* 1075 data , dan *Language* 765 data.

3. *Pre-Processing* Data

Pada tahap ini, peneliti merencanakan proses perancangan model sebagai berikut:

a. *Case folding*

Pada tahap ini, setiap kata dalam teks akan diubah menjadi huruf kecil menggunakan fungsi seperti lower() dalam *Python*. Ini mengurangi ketidakcocokan akibat perbedaan kapitalisasi antara kata yang serupa terutama huruf atau karakter alphabet menjadi huruf kecil.

b. *Text cleaning*

Tahapan *text cleaning* adalah tahap untuk penghapusan tanda baca yang tidak relevan, penghapusan angka yang tidak memberikan informasi penting untuk klasifikasi, dan spasi berlebih untuk meningkatkan kualitas teks.

c. *Tokenization*

Tokenisasi merupakan proses pemecahan teks menjadi unit-unit kecil yang disebut token, yang bisa berupa kata, frasa, atau bahkan karakter. Tokenisasi memudahkan model untuk memproses teks, karena model bekerja dengan token, bukan teks mentah.

4. Perancangan Arsitektur Model

Dalam tahap ini, akan direncanakan langkah-langkah proses perancangan arsitektur model secara rinci untuk memastikan setiap komponen dapat saling terintegrasi dengan baik.

a. Data Augmentation (cGAN)

Pada tahap ini, data dari hasil *pre-processing* akan diolah cGAN untuk menghasilkan data sintetik guna memperkaya dataset yang ada, sehingga hasil akhirnya tiap label memiliki jumlah dataset yang lebih seimbang.

Dalam penelitian ini, augmentasi cGAN tidak dilakukan pada bidang ilmu dengan jumlah data tertinggi yaitu *Economics* dengan jumlah 1142. Hal ini dikarenakan jumlah data tersebut sudah cukup efektif untuk digunakan dalam proses *training* BERT, sehingga proses augmentasi data sintetik hanya akan *diimplementasikan* pada 4 bidang ilmu lainnya yaitu *Education, Humanities*, *Social Science*, dan *Language.*

b. Text Embedding (BERT)

Proses Text Embedding dengan BERT melibatkan langkah-langkah yang memungkinkan model menghasilkan representasi semantik dalam bentuk vektor numerik (embedding) dari teks input. Representasi ini menangkap makna kata dan hubungan antar kata dalam konteks kalimat.

c. Model *Training* (*Clasissier*) dengan IndoBERT

Proses pelatihan model klasifikasi menggunakan IndoBERT dilakukan dengan memanfaatkan *pre-trained* model indobenchmark/indobert-base-p1 yang telah dilatih sebelumnya menggunakan korpus bahasa Indonesia. Model ini kemudian di-*fine-tune* untuk tugas klasifikasi topik jurnal ilmiah berdasarkan input teks berupa judul dan abstrak artikel.

5. Evaluasi Model

Ketika sistem selesai dibangun, dilakukan penilaian menyeluruh untuk menjamin bahwa sistem berfungsi dengan baik dan mencapai sasaran penelitian. Evaluasi dimulai dengan menguji kinerja model dalam sistem menggunakan masukan dari pengguna. Langkah ini ditujukan untuk mengevaluasi apakah Solusi yang dihasilkan relevan dan sesuai dalam mengatasi masalah yang ada, tidak hanya berjalan secara teknis, tetapi juga mampu menhasilkan prediksi yang relevan, akurat, dan sesuai dengan tujuan penelitian.

Setelah model selesai dilatih, metrik evaluasi diperlukan untuk mengetahui seberapa akurat model memprediksi topik artikel ilmihah. Metrik ini berdasarkan pada perbandingan antara hasil prediksi model dengan data yang sebenarnya. Metrik evaluasi yang digunakan pada penelitian ini adalah akurasi (*accuracy*), presisi (*precision*), *recall* (*sensitivity*), dan f1-*score*.

D. HASIL DAN PEMBAHASAN

1. Hasil Implementasi Sistem

Penelitian ini menghasilkan sebuah sistem klasifikasi bidang ilmu artikel ilmiah terindeks Garuda menggunakan metode cGAN – BERT. Proses penelitian diawali dengan mengumpulkan dataset

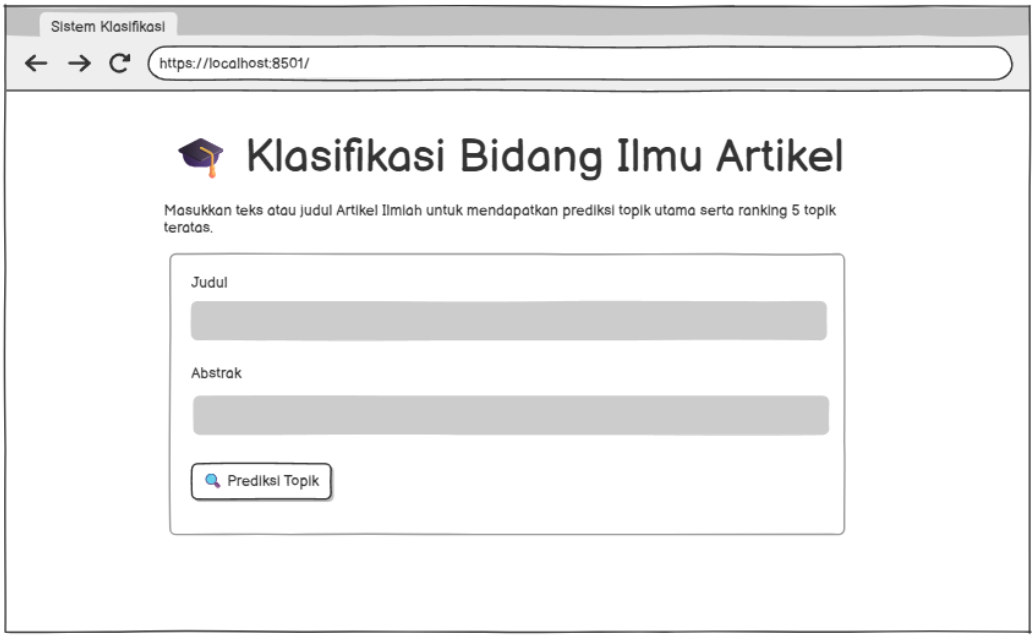
yang bersumber dari *platform* Garuda untuk diambil judul dan abstrak dari artikel ilmiah dengan 5 bidang ilmu yaitu *Economics*, *Education, Humanities*, *Social Science*, dan *Language*.

Setelah dilakukan pembersihan dan standarisasi data, ditemukan adanya ketimpangan distribusi antar bidang ilmu yaitu dengan jumlah keseluruhan 4340 data dengan rincian *Economics* 1142 data, *Public Health* 954 data, *Social Science* 824 data, *Computer Science & IT* 790 data, *Education* 630 data yang berarti jumlah data tidak seimbang satu sama lain dan dapat mengakibatkan sistem tidak belajar dengan baik dan akan terjadi bias pada salah satu bidang ilmu dengan jumlah data yang lebih banyak.

Untuk mengatasi ketidakseimbangan tersebut, diterapkan metode cGAN guna menghasilkan data sintetik yang menyerupai data asli. Proses training cGAN dilakukan selama 5000 *epoch* dengan menggunakan representasi fitur dari TF-IDF. Selanjutnya, seluruh data (asli dan sintetik) digunakan untuk melatih model IndoBERT (*indobenchmark/indobert-base*-p1) dalam melakukan klasifikasi. Pelatihan dilakukan selama 5 *epoch* menggunakan 80% data sebagai *training* dan 20% sebagai *validation*.

Setelah model berhasil dijalankan, langkah selanjutnya adalah mengimplementasikan aplikasi dalam bentuk *website*. *Website* ini dirancang agar pengguna dapat langsung memasukkan teks judul dan abstrak artikel ilmiah yang akan diteliti, kemudian sistem akan memproses *input* tersebut dan menampilkan hasil prediksi bidang ilmu pada halaman hasil. Dengan memanfaatkan *Streamlit*, aplikasi ini dapat dijalankan secara lokal atau diunggah ke server, sehingga dapat diakses dengan lebih mudah oleh pengguna lain.

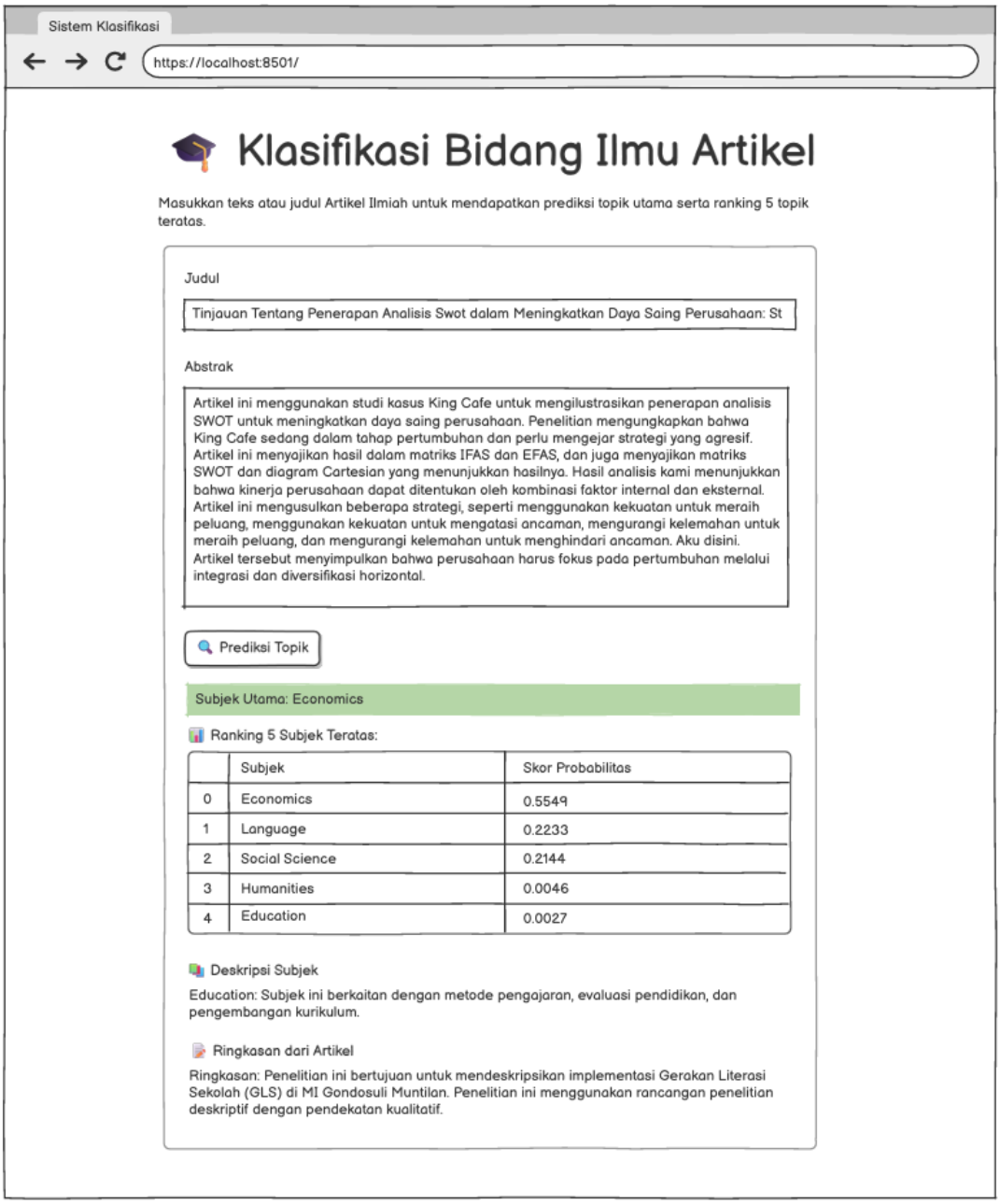
a. Hasil Perancangan *User Interface*



Gambar 3. Tampilan awal website

Gambar 3 merupakan halaman tampilan utama dari sistem klasifikasi artikel ilmiah terindeks Garuda ketika pengguna awal mengakses *website* tersebut. Pada halaman ini pengguna akan melihat tampilan *title* bertuliskan "Klasifikasi Bidang Ilmu Artikel", yang dibawahnya terdapat teks instruksi yang memerintahkan untuk memasukkan judul dan abstrak dari artikel yang akan di klasifikasikan, kemudian dibawahnya terdapat tombol "Prediksi Topik" dimana data akan dikirimkan *input* ke sistem agar dapat diproses dan menghasilkan *output* berupa hasil prediksi bidang ilmu yang paling relevan dengan data yang dimasukkan.

P- ISSN: 2355-6684, E-ISSN: 2776-6640
https://ojs.unpkediri.ac.id/index.php/noe



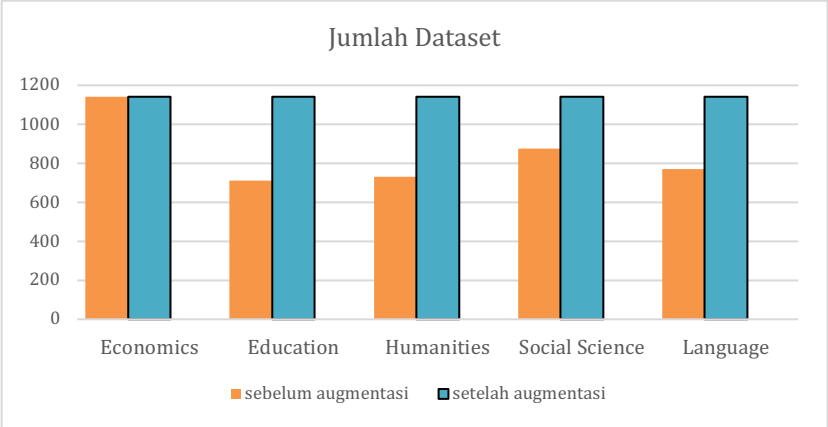
Gambar 4. Tampilan Hasil Prediksi Sistem

Gambar 4 merupakan halaman sistem klasifikasi yang menampilkan hasil dari proses prediksi berdasarkan judul dan abstrak artikel ilmiah yang dimasukkan oleh pengguna. Terdapat hasil prediksi subjek utama yang paling relevan dengan *input* teks, kemudian terdapat tabel yang menampilkan *ranking* 5 subjek teratas dengan 2 kolom yaitu kolom subjek dan kolom skor probabilitas. Data akan diurutkan dari skor tertinggi dan akan menjadi subjek utama yang berada di nomor 1 dan seterusnya hingga ke nomor 5 dengan bidang ilmu yang memiliki skor probabilitas paling kecil. Kemudian terdapat kolom "Deskripsi Subjek" yang berisikan deskripsi singkat artikel, juga terdapat kolom "Ringkasan dari Artikel" yang berisi ringkasan dari *input*an artikel.

b. Hasil cGAN – BERT

Hasil implementasi *Conditional Generative Adversarial Network* (cGAN) dan *Bidirectional Encoder Representations from Transformers* (BERT) dalam penelitian ini menunjukkan bahwa model mampu memberikan prediksi subjek berdasarkan judul dan abstrak dari artikel ilmiah yang dimasukkan oleh pengguna. Model bekerja dengan memanfaatkan pembuatan data sintetik pada proses augmentasi dengan cGAN dan pelatihan *fine-tuning* dengan BERT. Data sintetik yang berhasil ditambahkan sejumlah 1484 data sehingga total seluruh data menjadi 5710 data.

P- ISSN: 2355-6684, E-ISSN: 2776-6640
https://ojs.unpkediri.ac.id/index.php/noe



Gambar 5. *Chart* dataset sebelum dan sesudah augmentasi data

Dari gambar 5 diketahui bahwa proses augmentasi menggunakan cGAN menunjukan statistik data sebelum dan setelah proses augmentasi menghasilkan perbedaan jumlah data. Data sebelum augmentasi untuk tiap bidang ilmu adalah *Economics* 1142 data, *Education* 711 data, *Humanities* 730 data, *Social Science* 874 data, dan *Language* 769 data. Kemudian setelah melalu proses augmentasi data, jumlahnya menjadi lebih seimbang untuk ke-lima bidang ilmu yang tersedia yaitu menjadi 5710 data dengan rincian *Economics* 1142 data, *Education* 1142 data, *Humanities* 1142 data, *Social Science* 1142 data, dan *Language* 1142 data.

Hasil di atas menunjukkan bahwa penggabungan metode *Conditional Generative Adversarial* Network (cGAN) dan *Bidirectional Encoder Representations from Transformers* (BERT) berhasil menciptakan sistem klasifikasi artikel ilmiah terindeks garuda yang tepat dan relevan. Proses augmentasi data melalui cGAN dapat menyeimbangkan distribusi data pada setiap bidang ilmu dengan menciptakan artikel sintetik yang menyerupai dengan data asli dalam distribusi semantik. Model BERT yang dilatih ulang dengan data augmentasi menunjukkan kinerja yang baik dalam mengklasifikasikan artikel berdasarkan subjek utama seperti *Economics*, *Education, Humanities*, *Social Science*, dan *Language*. Hasil prediksi dari model menunjukkan pemahaman yang kontekstual mengenai judul dan abstrak artikel, dengan probabilitas klasifikasi yang konsisten terhadap isi yang tersedia.

Secara keseluruhan, penelitian ini membuktikan bahwa metode cGAN – BERT dapat menciptakan sistem klasifikasi artikel ilmiah yang tidak hanya tepat, tetapi juga responsif terhadap variasi subjek di *platform* Garuda. Sistem yang dibuat memiliki peluang untuk ditingkatkan lebih lanjut dalam skala yang lebih besar sebagai bagian dari sistem pencarian dan saran artikel ilmiah berbasis NLP.

2. Hasil Evaluasi Sistem

Evaluasi sistem klasifikasi berguna untuk mengukur performa model BERT setelah proses pelatihan dengan data hasil augmentasi menggunakan cGAN. Evaluasi penting dilakukan untuk menentukan sejauh mana model mampu melakukan tugas klasifikasi dengan benar. Dalam penelitian ini, evaluasi dilakukan menggunakan metrik evaluasi kinerja model seperti akurasi, presisi, *recall*, dan f1-*score*.

Tabel 1. Hasil Evaluasi Sistem

Subject	Precision	Recall	F1-Score	Support
Economics	0.79	0.79	0.79	229
Education	0.91	**0.96**	**0.94**	228
Humanities	0.81	0.79	0.80	228
Language	**0.93**	0.93	0.93	229
Social Science	0.89	0.86	0.87	228
Average				
Accuracy			0.87	1142
Macro avg			0.87	1142
Weighted avg			0.87	1142

Tabel 1 menunjukkan hasil pengujian sistem. Berdasarkan evaluasi menggunakan precision, recall, dan f1-score, kinerja model bervariasi di setiap bidang ilmu namun secara umum cukup baik. Economics

yang tidak mengalami augmentasi meraih skor 79% di semua metrik, namun masih terdapat false negative sehingga perlu variasi data. Education menunjukkan peningkatan signifikan setelah augmentasi dengan skor 91% di semua metrik, membuktikan bahwa data tambahan berhasil memperkuat representasi kelas ini. Humanities mencatat skor terendah (precision 81%, recall 79%, f1-score 80%) karena kesamaan kata kunci dengan bidang lain membuat model lebih sering keliru. Language mencapai skor tertinggi (93% di semua metrik) berkat keberagaman data augmentasi dan pola bahasa yang konsisten. Social Science memperoleh precision 89%, recall 86%, dan f1-score 87%, menunjukkan hasil baik meskipun masih perlu peningkatan dalam membedakan dari kategori lain.

Evaluasi pada 1.142 data uji menunjukkan akurasi model, *macro,* dan *weighted average* masing-masing 87%. Hasil ini membuktikan bahwa augmentasi cGAN dan embedding BERT berhasil meningkatkan distribusi kelas minoritas dan kinerja sistem klasifikasi artikel ilmiah Garuda secara relevan.

Secara keseluruhan, hasil evaluasi ini menegaskan bahwa strategi augmentasi data memberikan dampak positif terhadap sebagian besar kategori, terutama bidang dengan jumlah data awal yang terbatas. Meski demikian, beberapa bidang ilmu masih memerlukan penyesuaian strategi augmentasi untuk mengurangi kesalahan klasifikasi yang disebabkan oleh tumpang tindih topik dan kemiripan istilah. Optimalisasi ini diharapkan dapat meningkatkan performa model secara menyeluruh serta memastikan sistem mampu melakukan klasifikasi dengan akurasi yang konsisten di berbagai skenario data.

E. KESIMPULAN DAN SARAN

Berdasarkan hasil penelitian ini, dapat diambil kesimpulan bahwa sistem klasifikasi bidang ilmu artikel ilmiah terindeks Garuda dapat berjalan dengan baik, dengan didukung teknik augmentasi data menggunakan *Conditional Generative Adversarial Network* (cGAN) dan teknik *text embeding* untuk memahami konteks teks menggunakan *Bidirectional Encoder Representations from Transformers* (BERT). Permasalahan ketidak seimbangan data antar subjek berhasil di atasi dengan melakukan augmentasi pada subjek-subjek minoritas, seperti education, humanities, social science, dan language, sementara subjek economics dikecualikan karena jumlah data dianggap telah mencukupi.

Pendekatan ini memungkinkan sistem untuk menghasilkan distribusi data yang lebih seimbang. Ini menghasilkan proses klasifikasi yang lebih stabil dan tidak lagi bias terhadap topik mayoritas. Hasil evaluasi menunjukkan bahwa metrik klasifikasi seperti akurasi, presisi, *recall*, dan f1-*score* dapat ditingkatkan dengan menggunakan data hasil peningkatan, terutama berlaku untuk bidang ilmu yang sebelumnya memiliki data yang tidak seimbang. Selain itu, model BERT sangat baik dalam memahami konteks teks jurnal dan mengklasifikasikannya ke dalam subjek yang relevan. Oleh karena itu, integrasi antara augmentasi berbasis cGAN dan model BERT terbukti efektif dalam meningkatkan akurasi sistem klasifikasi bidang ilmu artikel ilmiah dan menyelesaikan masalah yang sering terjadi dalam pemrosesan data tidak seimbang yang berkaitan dengan tugas klasifikasi teks. Akan tetapi, masih terdapat kekurangan atau pun hal yang dapat diperbaiki pada penelitian selanjutnya. Adapun saran untuk penelitian selanjutnya ialah sebagai berikut:

1. Penggunaan metode augmentasi yang lebih beragam seperti back-translation, penggantian synonym kontekstual, atau paraphrasing model agar dapat menghindari duplikasi dan menghasilkan kalimat yang lebih natural dalam augmentasi berbasis GAN.
2. Perluasan cakupan bidang ilmu jurnal dan penelitian label multilabel, karena banyak artikel ilmiah yang mencakup lebih dari satu bidang, membuat model lebih fleksibel dalam pengklasifikasian artikel.
3. Banyak jurnal di Indonesia yang menggunakan campuran bahasa Inggris. Untuk menangani variasi bahasa ini dan meningkatkan kekuatan sistem klasifikasi, penelitian lanjutan dapat mencoba pendekatan multibahasa.

DAFTAR PUSTAKA

Adhikari, Ashutosh, Achyudh Ram, Raphael Tang, and Jimmy Lin. 2020. "DocBERT: BERT for Document Classification."

Anisatuzzumara. 2024. "Implementasi Latent Dirichlet Allocation (LDA) Dan K-Nearest Neighbors(KNN) Pada Sistem Rekomendasi Jurnal Terindeks GARUDA." *Αγαη* 15(1):37–48.

Bhat, Ranjith, and Raghu Nanjundegowda. 2025. "A Review on Comparative Analysis of Generative Adversarial Networks' Architectures and Applications." Journal of Robotics and Control (JRC) 6(1):53–64. doi: 10.18196/jrc.v6i1.24160.

Croce, Danilo, Giuseppe Castellucci, and Roberto Basili. 2020. "GAN-BERT: Generative Adversarial Learning for Robust Text Classification with a Bunch of Labeled Examples." Proceedings of the Annual Meeting of the Association for Computational Linguistics 2114–19. doi: 10.18653/v1/2020.acl-main.191.

Dameani, Tiara. 2021. "Analisis Panel Data Web Scraping Artikel Kekerasan Dalam Rumah Tangga Tahun 2019- 2020 Di DKI Jakarta." Jurnal Teknologi Informasi 7(1):43–49. doi: 10.52643/jti.v7i1.1321.

Hafiz, Y. A., and Endah Sudarmilah. 2023. "Implementasi Web Scraping Pada Portal Berita Online." Inisiasi 55–60. doi: 10.59344/inisiasi.v12i1.120.

Hajkowicz, Stefan, Conrad Sanderson, Sarvnaz Karimi, Alexandra Bratanova, and Claire Naughtin. 2023. "Artificial Intelligence Adoption in the Physical Sciences, Natural Sciences, Life Sciences, Social Sciences and the Arts and Humanities: A Bibliometric Analysis of Research Publications from 1960-2021." *Technology in Society* 74. doi: 10.1016/j.techsoc.2023.102260.

Islam, Saidul, Hanae Elmekki, Ahmed Elsebai, Jamal Bentahar, Najat Drawel, Gaith Rjoub, Witold Pedrycz, Computer Science, Abu Dhabi, and Saudi Arabia. 2023. "A C OMPREHENSIVE S URVEY ON A PPLICATIONS OF."

Kuang, Kun, Yuxiao Lin, Yuxian Meng, Xiaofei Sun, Qinghong Han, Jiwei Li, and Fei Wu. 2021. "BertGCN: Transductive Text Classification by Combining GCN and BERT." *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021* 1456–62. doi: 10.18653/v1/2021.findings-acl.126.

Li, Yang, Kangbo Liu, Ranjan Satapathy, Suhang Wang, and Erik Cambria. 2024. "Recent Developments in Recommender Systems: A Survey [Review Article]." *IEEE Computational* Intelligence *Magazine* 19(2):78–95. doi: 10.1109/MCI.2024.3363984.

Ma, Lijing, and Shiru Qu. 2022. "Application of Conditional Generative Adversarial Network To." (March 2021).

Nayla, Adine, Casi Setianingsih, and Burhanuddin Dirgantoro. 2023. "Deteksi Hate Speech Pada Twitter." *EProceeding of Engineering* 10(1):256.

Ribas, Lucas C., Wallace Casaca, and Ricardo T. Fares. 2025. "Conditional Generative Adversarial Networks and Deep Learning Data Augmentation: A Multi-Perspective Data-Driven Survey Across Multiple Application Fields and Classification Architectures." *AI (Switzerland)* 6(2). doi: 10.3390/ai6020032.

Rogers, Anna, Olga Kovaleva, and Anna Rumshisky. 2020. "A Primer in Bertology: What We Know about How Bert Works." *Transactions of the Association for Computational Linguistics* 8:842–66. doi: 10.1162/tacl_a_00349.

Sa'adah, Farikhatus. 2022. "Klasifikasi Bidang Ilmu Pada Publikasi Terindeks Garuda Menggunakan Metode K-Nearest Neighbor (K-Nn)." *Angewandte Chemie International Edition, 6(11), 951–952.* 5(2):5–24.

Shah, Momna Ali, Muhammad Javed Iqbal, Neelum Noreen, and Iftikhar Ahmed. 2023. "An Automated Text Document Classification Framework Using BERT." *International Journal of Advanced Computer Science and Applications* 14(3):279–85. doi: 10.14569/IJACSA.2023.0140332.

Supardi, Cholid Fajar. 2023. "Final Project Trend Search System Final Project Title of Unissula Informatics Engineering Students Using Keyword Extraction." Universitas Islam Sultan Agung.

Suprapti, Tati, Dian Ade Kurnia, Doni Anggara, Rananda Deva Rian, and Aldi Setiawan. 2023. "Implementasi Model Algoritma Generative Adversarial Network (Gan) Pada Sistem Presensi Berbasis Deteksi Wajah (SIDEWA)." *Tematik* 9(2):231–36. doi: 10.38204/tematik.v9i2.1048.

Syarifudin, Faisal. 2022. "Klasifikasi Artikel-Artikel Jurnal Pustakaloka Berdasarkan Skema Jita." Fihris*: Jurnal Ilmu Perpustakaan Dan Informasi* 17(1):20. doi: 10.14421/fhrs.2022.171.20-37.

Wang, Zhengwei, Qi She, and Tomás E. Ward. 2021. "Generative Adversarial Networks in Computer Vision: A Survey and Taxonomy." *ACM Computing Surveys* 54(2). doi: 10.1145/3439723.

Widiansyah, Muhammad, Fathia Frazna Az-zahra, and Agung Pambudi. 2021. "Fine-Tuning Model Indobert (Indonesian Bidirectional Encoder Representations from Transformers) Untuk Analisis Sentimen Berbasis Aspek Pada Aplikasi M-Paspor."

Wijaya, Bhianta, and Edi Surya Negara. 2022. "Penerapan Garuda Smart City Model Dalam Menganalisa Kesiapan Pemerintah Kabupaten Tulang Bawang Barat Dalam Membangun Konsep Smart City." *CogITo Smart Journal* 8(2):524–36. doi: 10.31154/cogito.v8i2.436.524-536.