BAB III
ANALISIS DAN PERANCANGAN

3.1 Analisis

Proses analisis bertujuan untuk memahami permasalahan yang ada dan menyusun kebutuhan sistem chatbot berbasis Natural Language Processing (NLP) untuk mendeteksi dan menanggulangi cyberbullying.

3.1.1 Identifikasi Masalah

Cyberbullying di platform digital seperti Discord semakin meningkat dan menimbulkan berbagai masalah serius antara lain:

1. Peningkatan Kasus: Jumlah kasus cyberbullying terus bertambah, terutama di kalangan remaja
2. Kesulitan Deteksi: Penggunaan bahasa sandi, sarkasme, atau konteks spesifik menyulitkan deteksi cyberbullying.
3. Anonimitas Pelaku: Kemudahan membuat akun anonim menyulitkan identifikasi dan penanganan pelaku.
4. Kurangnya Pengawasan: Banyak grup atau channel kekurangan pengawasan efektif.
5. Variasi Bentuk Bullying: Cyberbullying muncul dalam berbagai bentuk, menyulitkan identifikasi dengan satu pendekatan.
6. Keterbatasan Moderasi Manual: Moderasi manual seringkali tidak efektif untuk volume pesan yang besar.
7. Kesenjangan Pemahaman: Banyak pengguna mungkin tidak menyadari perilaku cyberbullying atau cara melaporkannya.
8. Keterlambatan Respons: Tindakan terhadap cyberbullying sering terlambat diambil, memperpanjang dampak negatif pada korban.

3.1.2 Pemecahan Masalah

Pemecahan masalah dalam permasalahan diatas merupakan langkah krusial untuk meningkatkan kinerja dan efektivitas chatbot. Berikut adalah beberapa cara yang dapat diterapkan untuk mengatasi masalah yang telah diidentifikasi sebelumnya:

1. Deteksi Real-time

Implementasi model IndoBERT yang disesuaikan untuk mendeteksi berbagai bentuk cyberbullying secara real-time.

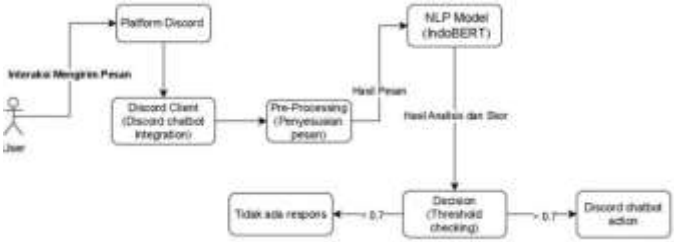
2. Respons Cepat dan Otomatis
   Chatbot memberikan peringatan otomatis kepada pelaku potensial dan memberikan notifikasi instan kepada moderator atau admin grup untuk tindakan lebih lanjut.
3. Edukasi Pengguna
   Chatbot menyediakan informasi edukatif tentang cyberbullying dan dampaknya, juga memberikan panduan langkah-langkah untuk melaporkan atau menangani cyberbullying.
4. Analisis Sentiment dan Konteks
   Penggunaan BERT untuk analisis sentimen mendalam, memahami sarkasme dan bahasa tersirat.
5. Sistem Pelaporan Terintegrasi
   Fitur pelaporan mudah digunakan yang terintegrasi dengan chatbot.
6. Anonimisasi dan Perlindungan Privasi
   Implementasi teknik anonimisasi data untuk melindungi identitas pelapor.
7. Adaptabilitas Model
   Pembaruan berkala model IndoBERT dengan data baru untuk mengatasi evolusi bahasa cyberbullying.
8. Integrasi multi-platform
   Pengembangan antarmuka yang kompatibel dengan berbagai platform seperti Discord.

3.2 Perancangan

Tahap perancangan adalah langkah penting untuk memastikan bahwa sistem chatbot berfungsi dengan baik sesuai kebutuhan. Pada tahap ini, berbagai komponen dan mekanisme sistem dirancang, termasuk alur kerja chatbot, struktur data, desain antarmuka pengguna, serta strategi pengujian. Perancangan dilakukan untuk memastikan bahwa chatbot dapat mendeteksi dan menanggulangi cyberbullying secara efektif di platform Discord.

3.2.1 Perancangan Sistem

Perancangan sistem merupakan tahap penting untuk memastikan solusi chatbot dapat berjalan dengan baik dan sesuai dengan kebutuhan. Sistem chatbot anti-cyberbullying dirancang khusus untuk platform Discord, menggunakan model BERT yang sudah disesuaikan untuk deteksi cyberbullying.



Gambar 3.2.1 Perancangan Sistem

Gambar diatas menjelaskan tentang cara kerja chatbot menggunakan model IndoBERT. Sistem ini bekerja dengan mengintegrasikan Discord Bot dengan model NLP berbasis IndoBERT yang telah dilatih khusus untuk memahami bahasa Indonesia, termasuk slang, singkatan, dan kata-kata kasar yang sering muncul di platform media sosial.

Bot ini berperan sebagai moderator otomatis di dalam server Discord, memindai setiap pesan yang dikirimkan oleh pengguna, dan melakukan analisis untuk menentukan apakah pesan tersebut mengandung unsur cyberbullying, ujaran kebencian, atau potensi kekerasan verbal. Sistem melakukan klasifikasi pada pesan yang masuk dan memberikan label tertentu berdasarkan hasil analisis model. Apabila terdeteksi pesan yang memenuhi kriteria sebagai pesan bullying, bot memberikan tindakan preventif sesuai dengan kebijakan yang telah ditetapkan, seperti memberikan peringatan, menghapus pesan, atau melakukan mute pada pengguna.

Dengan adanya sistem ini, diharapkan moderasi pada server Discord dapat berjalan secara otomatis, sehingga mengurangi beban

moderator manual dalam memantau obrolan, serta menciptakan lingkungan digital yang lebih aman dan nyaman bagi pengguna, terutama remaja dan anak-anak yang sering menjadi target cyberbullying.

Proses Kerja sistem diatas adalah sebagai berikut:

1. User mengirimkan pesan di dalam server yang ada di platform Discord
2. Discord Bot yang sudah di integrasikan dengan Model dan sudah berada di dalam server menerima pesan tersebut
3. Pesan di sesuaikan dan hasil pelabelan pesan di terima oleh model IndoBERT
4. Model memberikan skor yang sesuai dengan hasil Analisa model
5. Bot beraksi sesuai hasil skor threshold yang sudah didetapkan dan disesuaikan

Untuk mendukung fungsionalitas tersebut, berikut adalah komponen utama darisistem chatbot yang dikembangkan:

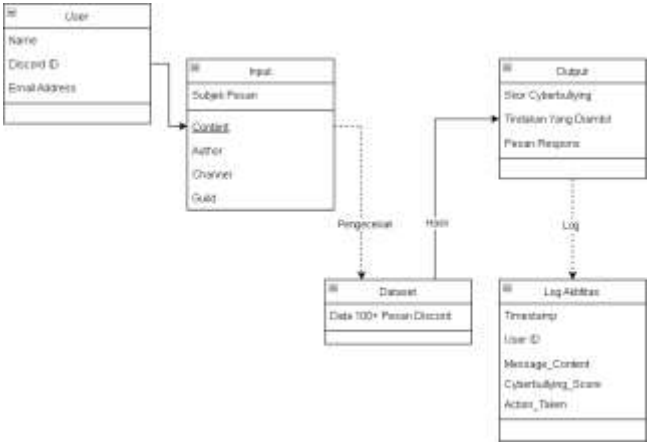
1. Modul Discord Client
   Menggunakan Discord.py library untuk koneksi ke Discord API.Menangani event pesan masuk dari server Discord.
2. Modul Pre-Processing
   Membersihkan pesan (menghapus karakter khusus, mengubah ke lowercase).
   Memfilter pesan untuk menghindari pemrosesan pesan bot atau command.
3. Model IndoBERT
   Model IndoBERT pra-terlatih yang di-fine-tune untuk klasifikasi cyberbullying.
   Input: Pesan yang telah di-pre-process.
   Output: Skor probabilitas cyberbullying (0-1)
4. Modul Keputusan
   Menggunakan threshold (misalnya 0.7) untuk menentukan status cyberbullying.

Menentukan tindakan berdasarkan skor (peringatan, mute, atau notifikasi moderator).

5. Modul Respons
   Menyimpan template respons untuk berbagai tingkat cyberbullying. Memilih dan memformat respons yang sesuai.
6. Modul Aksi Discord
   Mengirim pesan peringatan ke channel atau DM. Melakukan tindakan moderasi seperti mute jika diperlukan.

3.2.2 Perancangan Data

Pada tahap ini, perancangan data difokuskan pada penentuan struktur dan kebutuhan informasi yang diperlukan oleh chatbot. Proses ini mencakup identifikasi tipe data yang relevan, sumber pengumpulan data, metode penyimpanan, serta bagaimana data tersebut saling terhubung.



Gambar 3.2.2 Perancangan Data

Gambar diatas menjelaskan tentang data yang digunakan dalam pengembangan chatbot. Dalam konteks chatbot yang bertugas mendeteksi dan mengatasi cyberbullying di platform seperti Discord, desain data diatas memiliki peran penting untuk memastikan chatbot mampu mengenali pesan berbahaya dan memberikan respons yang tepat. Penjelasan tentang isi data sebagai berikut:

1. User**:** User mewakili pengguna yang berkomunikasi dengan cara memberikan pesan.
2. Input**:** Berisi subjek pesan dari pengguna dan detail informasi konten, pengirim, platform komunikasi (channel), dan grup (guild).
3. Dataset**:** Berisi data pesan teks cyberbullying maupun non-cyberbullying
4. Output**:** Hasil dari respons dataset yang terdiri dari skor cyberbullying, tindakan yang diambil, dan pesan respons
5. Log Aktifitas**:** Log aktifitas dari detail informasi yang dijalankan oleh chatbot.

3.2.3 Perancangan User Interface

Gambar 3.2.3 Perancangan User Interface

3.2.4 Perancangan Dataset

Dalam pengembangan chatbot anti-cyberbullying berbasis NLP ini, dataset menjadi komponen penting untuk melatih dan menguji model agar mampu mendeteksi pesan yang mengandung unsur cyberbullying dengan

akurat. Dataset yang digunakan dirancang untuk mencakup berbagai variasi teks yang sering muncul di media sosial, khususnya pada platform Discord, dengan fokus pada bahasa Indonesia, termasuk bahasa gaul, singkatan, serta kata-kata kasar yang umum digunakan.

Dataset yang dirancang memiliki beberapa label kategori, antara lain:

1. Cyberbullying: Pesan yang mengandung unsur perundungan, penghinaan, pelecehan, atau ancaman secara verbal.
2. Non-Cyberbullying: Pesan yang tidak mengandung unsur perundungan dan dianggap aman.

Sumber dataset diperoleh dari:

1. Cyberbullying Dataset yang diambil dari HuggingFace aditdwi123/cyber-bullying-dataset.
2. Hate Speech Dataset yang diambil dari Github okkyibrohim/id-multi-label-hate-speech-and-abusive-language-detection.

Langkah-langkah perancangan dataset adalah sebagai berikut:

1. Pengumpulan Data: Mengumpulkan data pesan teks dari dataset publik dan hasil simulasi percakapan.
2. Preprocessing Data: Melakukan pembersihan data dengan menghapus URL, emoji jika tidak relevan, dan normalisasi teks untuk mengurangi noise pada data.
3. Pelabelan Data: Melakukan pelabelan secara manual dengan membaca konteks setiap pesan dan menentukan label sesuai kategori yang telah ditentukan.
4. Balancing Data: Mengatur jumlah data pada setiap label agar seimbang untuk menghindari bias model selama pelatihan.

5. Format Dataset: Dataset disusun dalam format CSV dengan kolom *id*, *kalimat*, dan *sentimen*, agar mudah digunakan pada pipeline fine-tuning IndoBERT.

6. Split Data: Dataset dibagi menjadi *train* dan *test set* dengan proporsi 80%:15% atau 90%:10% untuk memastikan evaluasi model yang optimal.

Perancangan dataset ini dirancang agar model dapat mengenali pola kata, frasa, serta konteks yang dapat dikategorikan sebagai cyberbullying pada percakapan di Discord, sehingga chatbot dapat memberikan deteksi yang akurat dan dapat melakukan mitigasi secara otomatis.

3.2.5 Perancangan Model

Pada tahap ini dilakukan perancangan model deteksi cyberbullying menggunakan teknologi Natural Language Processing (NLP) dengan memanfaatkan model IndoBERT, yang merupakan model BERT yang dilatih khusus untuk Bahasa Indonesia. Model ini digunakan sebagai inti dari sistem pendeteksi pesan yang mengandung unsur cyberbullying pada platform Discord.

1. Pemilihan Model

Langkah pertama adalah memilih model dasar yang digunakan. Dalam implementasi ini, digunakan model indobenchmark/indobert-base-p1 dari Hugging Face karena telah dilatih secara khusus pada korpus Bahasa Indonesia, termasuk teks formal dan informal. Ini membuatnya sangat cocok untuk memproses bahasa sehari-hari yang banyak digunakan di platform Discord, termasuk singkatan, bahasa gaul, dan struktur kalimat yang tidak baku.

IndoBERT memiliki arsitektur transformer berbasis BERT yang mampu memahami konteks kalimat secara dua arah (bidirectional), yang sangat penting untuk menangkap makna tersirat dalam kasus cyberbullying, seperti sarkasme atau ejekan halus.

2. Fine-Tuning Model

Setelah model dasar ditentukan, dilakukan proses fine-tuning menggunakan dua dataset berbahasa Indonesia yang relevan:

1. Cyberbullying dataset dari cyberbullying_v2.csv (HuggingFace: aditdwi123/cyber-bullying-dataset)
2. Hate speech dataset dari hatespeech.csv (Kaggle: okkyibrohim/id-multi-label-hate-speech-and-abusive-language-detection)

Total gabungan data mencapai lebih dari 10.000 kalimat, masing-masing dilabeli sebagai positif (non-cyberbullying) atau negatif (cyberbullying). Label dikonversi menjadi format biner (0 = positif, 1 = negatif).

Fine-tuning dilakukan dengan skema sebagai berikut:

1. Preprocessing data: Meliputi pemetaan label, penggabungan dataset, dan tokenisasi menggunakan tokenizer indobenchmark/indobert-base-p1. Panjang maksimum tokenisasi diatur pada 128 token.
2. Pemisahan Dataset: Dataset pengujian pertama digabung dan dibagi menjadi 80% data pelatihan dan 20% data pengujian, Dataset pengujian kedua digabung dan dibagi menjadi 90% data pelatihan dan 10% data pengujian menggunakan train_test_split() dengan seed=42.
3. Pelatihan Model:
    1. Epoch: 5
    2. Batch Size: 8
    3. Optimizer: AdamW
    4. Strategi Evaluasi dan Penyimpanan: Setiap Epoch

5. Model disimpan secara otomatis berdasarkan nilai F1-score terbaik (load_best_model_at_end=True)

4. Framework: Digunakan framework Hugging Face Trainer untuk menangani proses pelatihan dan evaluasi model.

3. Evaluasi Model

Setelah proses pelatihan, model dievaluasi menggunakan metrik berikut:

1. Accuracy: Mengukur ketepatan klasifikasi secara keseluruhan.

2. F1-Score (macro average): Digunakan untuk menangani ketidakseimbangan kelas, memberikan bobot yang setara pada masing masing kelas

3. Visualisasi: Plot hasil accuracy dan F1-score per epoch disimpan dalam bentuk grafik PNG

4. Export Hasil: Metrik hasil pelatihan disimpan ke dalam file CSV (eval_metrics_epoch5.csv), dan model beserta tokenizer disimpan dalam folder terpisah untuk digunakan pada integrasi dengan bot discord

Evaluasi menunjukkan performa model stabil selama proses pelatihan dengan tren kenaikan pada metrik evaluasi. Confusion matrix dan threshold prediksi dapat ditentukan pada tahap integrasi selanjutnya sesuai kebutuhan sistem.

3.3 Rancangan Pengujian

Pengujian sistem chatbot untuk mendeteksi dan menanggulangi cyberbullying di dalam grup bertujuan memastikan bahwa chatbot berfungsi efektif dalam lingkungan diskusi yang melibatkan banyak pengguna secara bersamaan. Pengujian dilakukan untuk menilai performa, ketepatan deteksi, dan pengalaman pengguna di dalam 1 channel server Discord.

3.3.1 Metodologi Pengujian

Metodologi pengujian ini dirancang untuk mengevaluasi performa dan efektivitas chatbot anti-cyberbullying pada platform Discord. Chatbot diuji melalui minimal 1 channel di server Discord, masing-masing dengan minimal 2 anggota, untuk memastikan chatbot dapat memberikan respons yang tepat, akurat, dan konsisten dalam berbagai skenario. Metodologi ini mencakup beberapa tahapan berikut:

1. Pengujian Black Box
   Menguji fungsi chatbot tanpa melihat kode sumber. Fokus pada interaksi pengguna dan chatbot untuk memverifikasi apakah respons dan perilaku sesuai dengan yang diharapkan.
2. Pengujian Respons Kontekstual
   Menguji kemampuan chatbot dalam memahami percakapan berurutan atau diskusi panjang yang melibatkan banyak partisipan. Pengujian ini fokus pada bagaimana chatbot mempertahankan konteks dalam memberikan respons.
3. Pengujian Model
   Menguji keakuratan dan hasil evaluasi model dengan dataset yang diberikan.

3.3.2 Skenario Pengujian

Skenario pengujian adalah situasi yang dirancang untuk mengevaluasi kemampuan dan kinerja chatbot dalam mendeteksi dan menanggulangi cyberbullying di platform Discord. Berikut adalah beberapa skenario yang dapat digunakan:

1. Pengujian Respons terhadap Pesan Bernada Kasar atau Ujaran Kebencian

   Pengujian respons terhadap pesan kasar atau ujaran kebencian bertujuan untuk menilai kemampuan chatbot dalam mendeteksi elemen cyberbullying dalam percakapan grup. Dalam skenario ini, anggota grup mengirimkan pesan yang mengandung kata-kata menyerang. Chatbot diharapkan dapat memberikan peringatan kepada pengirim atau menyarankan admin grup untuk mengambil tindakan, sehingga membantu menjaga lingkungan komunikasi yang aman di platform digital.
2. Pengujian Respons terhadap Percakapan Bertingkat (Multi-turn Conversation)

   Pengujian respons terhadap percakapan bertingkat bertujuan untuk memastikan chatbot dapat mengikuti dialog yang berlangsung dalam beberapa langkah dengan banyak pengguna. Dalam skenario ini, beberapa anggota grup berbicara secara bergantian, dan chatbot harus memahami konteks percakapan yang sedang dibahas. Chatbot diharapkan dapat memberikan respons yang relevan meskipun interaksi bersifat dinamis dan melibatkan beberapa partisipan.
3. Pengujian Terhadap Pesan dengan Bahasa Tidak Formal dan Singkatan

   Pengujian terhadap pesan yang menggunakan bahasa tidak formal dan singkatan bertujuan untuk menilai kemampuan chatbot dalam memahami bahasa sehari-hari yang sering digunakan dalam percakapan grup, termasuk singkatan dan emoji. Dalam skenario ini, pengguna mengirimkan pesan dengan bahasa gaul atau singkatan. Chatbot diharapkan dapat menginterpretasikan maksud pesan tersebut dan memberikan respons yang akurat.

3.3.3 Kriteria Keberhasilan

1. Akurasi Deteksi: Model dapat mempelajari dan mengklasifikasikan ujaran kebencian dengan tingkat akurasi minimal 80%.
2. Konsistensi: Respons tetap relevan di semua channel uji dan konsisten di berbagai konteks percakapan.
3. Konsistensi Respons di Grup: Respons harus tetap relevan dan sesuai, meskipun melibatkan banyak pengguna atau bahasa yang tidak formal.