NIM : 252012154
Nama : Muhammad Kurnia Wahidin
Kelas : TI - RPL
Mata Kuliah : Smart Data Intelligence and AI Systems

CHATBOT INTELIJEN UNTUK KONSULTASI TENTANG DATA TUNGGAL SOSIAL EKONOMI NASIONAL (DTSEN)

I. Deskripsi Permasalahan

A. Profil Instansi

Dinas Sosial Kabupaten Pacitan merupakan satuan organisasi perangkat daerah (OPD) di bawah Pemerintah Kabupaten Pacitan, Provinsi Jawa Timur, yang bertugas melaksanakan urusan pemerintahan di bidang sosial.

Secara umum, Dinas Sosial bertanggung jawab untuk:

1. Pelayanan dan Rehabilitasi Sosial: Menangani Penyandang Disabilitas, Lanjut Usia (Lansia), Anak yang memerlukan perlindungan khusus (Anak Terlantar, Anak Berhadapan dengan Hukum), Tuna Sosial, Korban Penyalahgunaan NAPZA, dll.
2. Perlindungan Jaminan Sosial: Penyaluran bantuan sosial (Bansos) dari pemerintah pusat/daerah, seperti Program Keluarga Harapan (PKH), Bantuan Pangan Non Tunai (BPNT), Bantuan Sosial Tunai (BST), dan program sejenis.
3. Pemberdayaan Sosial: Pemberdayaan fakir miskin, komunitas adat terpencil, serta peningkatan kemandirian kelompok rentan.
4. Penanganan Masalah Kesejahteraan Sosial: Penanganan gelandangan dan pengemis (Gepeng), serta penanggulangan kemiskinan dari aspek sosial.
5. Tanggap Darurat Bencana (dari aspek sosial): Pendataan, penyaluran bantuan, dan pendampingan korban bencana alam.

Layanan Publik yang Diberikan :

★ Pendataan dan verifikasi penerima bantuan sosial.
★ Pemberian surat keterangan (misal untuk disabilitas atau fakir miskin).
★ Layanan konsultasi dan pendampingan masalah sosial.
★ Pelayanan panti/rehabilitasi.

B. Identifikasi permasalahan

1. **Beban Kerja Petugas Melonjak**

   Banyaknya pertanyaan berulang dari masyarakat terkait status, verifikasi, dan proses DTSEN membuat petugas kewalahan menangani konsultasi dasar.

2. **Akses Informasi Terbatas**
   Masyarakat kesulitan mendapatkan informasi real-time tentang DTSEN (seperti status data, jadwal update, alur perbaikan data) di luar jam kerja.
3. **Pemahaman Masyarakat yang Terbatas**
   Banyak mispersepsi tentang tujuan DTSEN, syarat, dan manfaatnya, yang membutuhkan penjelasan konsisten dan mudah dipahami.
4. **Layanan Non-Stop**
   Kebutuhan layanan konsultasi 24/7 untuk pertanyaan mendesak, terutama saat pendistribusian bansos.
5. **Efisiensi Proses**
   Kurangnya alat bantu awal untuk pra-screening pertanyaan sebelum escalasi ke petugas, sehingga proses konsultasi manual lambat.
6. **Disinformasi**
   Penyebaran hoaks atau informasi tidak resmi tentang DTSEN di masyarakat yang butuh koreksi cepat dan otomatis.

II. Tren Teknologi & Relevansi Riset

Chatbot yang didesain oleh penulis ini tepat berada di tengah tren AI terkini, kebutuhan nyata pemerintah, serta riset pengembangan yang relevan.
Berikut detail nya:

A. Tren Teknologi Terkini yang diimplementasikan ke Chatbot ini adalah :

1. **RAG (Retrieval - Augmented Generation)** - Tren utama yang sering diimplementasikan pada 2024, yang dapat secara signifikan mengurangi halusinasi AI
2. **Vector Databases** - FAISS/Chroma/Pinecone untuk search semantik
3. **Small Language Models** - sebagai efisiensi jika dibandingkan dengan LLM yang menggunakan cukup banyak resource. Model yang digunakan adalah SBERT
4. **Domain-Specific AI** - chatbot dengan batasan yang sempit (secara khusus hanya melayani diskusi / konsultasi terkait DTSEN)
5. **On-premise AI** - Privasi data jika akan dikembangkan dengan sistem internal yang memiliki data sensitif (misal identitas penerima bantuan), dan mengurangi ketergantungan dengan cloud API.

B. Relevansi Riset

1. **Digitalisasi pemerintahan** - DTSEN sebagai bagian dari transformasi digital pemerintah.
2. **AI untuk public service** - meningkatkan akses informasi
3. **Hybrid approach** - AI dikombinasikan dengan aturan (role-based) untuk mamastikan proses sesuai dengan aturan yang berlaku. Misalnya mekanisme

usulan penerima bantuan, pendataan keluarga miskin, validasi data, dan sebagainya.

4. **Scalable architecture** - chatbot ini sangat memungkinkan untuk dikembangkan menjadi skala yang lebih besar.

III. Data Spasial & GeoAI

Integrasi dengan data spasial dan GeoAI, dalam rancang bangun chatbot intelijen memang tidak melibatkan secara khusus tentang data spasial. Namun pada prakteknya Data Tunggal Sosial Ekonomi Nasional (DTSEN) sebagai rujukan / dasar distribusi bantuan terdapat data location (lat-long) terutama untuk KK yang berada pada desil rendah (desil 1-2) yang diverifikasi secara berkala pada proses verifikasi dan validasi data.

Pada integrasi yang lebih lanjut chatbot intelijen ini akan terintegrasi dengan terintegrasi dengan solusi Vehicle Routing Problem (VRP) Efisiensi Biaya Distribusi Paket Bantuan Sosial yang telah penulis buat sebelumnya untuk memenuhi tugas UAS di Mata Kuliah Model dan Simulasi
(Script lengkap ada di repository : https://github.com/Kurnia-Wahidin/uas_model_simulasi)

IV. Big Data & Pengolahan Data

A. Identifikasi Jenis Data

1. **Data terstruktur** adalah data dalam format terstandart dalam bentuk record dan field.
   Dalam hal ini, Data Tunggal Sosial Ekonomi Nasional (DTSEN) adalah data terstruktur karena terdapat struktur NIK, Nama, Alamat, dst.
2. **Data Tidak Terstruktur** adalah data yang disajikan tanpa memiliki format yang baku (tanpa baris dan kolom). Data tidak terstruktur biasanya berupa teks bebas yang deskripsi panjang, file pdf dan docx, gambar, video-audio, dsb.
   Dalam hal ini data dokumen training (pdf) yang berisi tentang peraturan pemerintah tentang DTSEN, teks-teks yang berisi definisi tentang DTSEN, menjelaskan alur / SOP usulan ke DTSEN, dan sebagainya. File-file dokumen berupa data training tersebut berada di directory ./sop_documents dalam aplikasi chatbot ini
3. **Volume** , big bata dapat berupa data dengan volume yang besar secara harfiah. Ini menyebabkan teknik untuk menghandle data menjadi sedikit berbeda jika dibandingkan dengan data dengan volume kecil. Pemanfaatan index, parsing, sorting memudahkan dalam menghandle data dengan volume besar tersebut, meski tidak kalah penting juga spesifikasi hardware juga

sangat berpengaruh.
Dalam hal ini, ketersediaan data training yang besar akan sangat menentukan kualitas respon dari chatbot intelijen yang dibangun. Dan dengan data training yang besar akan berpengaruh juga terhadap kebutuhan / spesifikasi hardware yang memadai.

4. **Velocity**, salah satu karakteristik utama dalam big data yang mengacu pada seberapa cepat data dihasilkan, dikumpulkan, dan diproses secara *real-time* atau mendekati *real-time*. Ini mencakup kecepatan aliran data masuk dan kebutuhan pemrosesan instan untuk pengambilan keputusan.
   Dalam hal ini, seberapa cepat data training dapat diproses masih menjadi tantangan utama. Selain itu perubahan data di DTSEN sendiri menjadi sebuah keharusan, karena kondisi ekonomi setiap keluarga akan selalu berubah-ubah.
5. **Variety**, keberagaman jenis, format, dan sumber data yang dikumpulkan, mencakup data terstruktur (tabel), semi-terstruktur (JSON/XML), hingga tidak terstruktur (video, gambar, teks). Ini menuntut pendekatan pengolahan yang berbeda karena data berasal dari berbagai sumber.
   Dalam hal ini, data training yang saat ini dapat diproses memiliki beberapa versi diantaranya file text (*.txt), file pdf (*.pdf), file document (*.docx), dan akan terus dituntut berkembang untuk dapat memproses format-format data yang lain pada update selanjutnya.

B. Alur Pengolahan Data

1. Pengumpulan Data untuk *big data* adalah proses sistematis menghimpun volume informasi besar, cepat, dan beragam (terstruktur/tidak terstruktur) dari berbagai sumber—seperti IoT, media sosial, dan transaksi—untuk dianalisis guna wawasan bisnis. Teknik utamanya meliputi web scraping, sensor, streaming API, dan log file, yang bertujuan mendukung pengambilan keputusan strategis, efisiensi operasional, dan inovasi produk.
2. Pembersihan Data (*data cleansing*) untuk big data adalah proses krusial mengidentifikasi dan memperbaiki data mentah yang tidak konsisten, salah, atau tidak lengkap untuk meningkatkan kualitas, akurasi, dan kegunaan dalam analisis, terutama mengingat volume data yang masif. Langkah utamanya mencakup penghapusan duplikat, penanganan data hilang (*missing values*), koreksi format, dan standarisasi, sering kali menggunakan teknik otomatisasi dan AI karena kerumitannya.
   Langkah-langkah Utama Pembersihan Data :
   a) Menangani Data Hilang (Missing Data): Mengisi nilai kosong dengan mean, median, atau modus, atau menghapus baris/kolom jika jumlahnya tidak signifikan.

b) Menghapus Data Duplikat: Mengidentifikasi dan menghapus data yang redundan dari berbagai sumber data.
c) Standarisasi Format: Memastikan konsistensi format data, seperti penulisan tanggal, satuan, atau penggunaan kapitalisasi (misal: "electronics" dan "Electronics").
d) Memperbaiki Data Salah: Mengoreksi kesalahan ejaan, format sintaksis yang salah, atau nilai yang tidak masuk akal.
e) Data Relevan: Menghapus data yang tidak relevan atau tidak diperlukan untuk tujuan analisis spesifik.

3. Penyimpanan data untuk big data membutuhkan arsitektur terukur (skalabel) yang mampu menampung volume besar, kecepatan tinggi, dan variasi data (terstruktur/tidak terstruktur). Solusi utamanya melibatkan Data Lake (seperti Hadoop HDFS/S3), database NoSQL (MongoDB, Cassandra), serta penyimpanan berbasis objek atau cloud untuk fleksibilitas kapasitas hingga petabyte.
   a) Teknologi dan Solusi Penyimpanan Big Data Utama:
      (1) Data Lake (Hadoop HDFS): Menggunakan sistem file terdistribusi (HDFS) yang memungkinkan penyimpanan data dalam jumlah besar dan format apa pun, yang kemudian diproses menggunakan Hadoop MapReduce.
      (2) Penyimpanan Cloud (Cloud Storage): Solusi seperti AWS S3, Google Cloud Storage, atau Azure Blob Storage menawarkan penyimpanan hemat biaya dengan skalabilitas tak terbatas.
      (3) Database NoSQL: Dirancang untuk data tidak terstruktur atau semi-terstruktur, seperti MongoDB (dokumen), Cassandra (kolom), dan HBase.
      (4) Penyimpanan Berbasis Objek: Opsi ini (seperti Cloudian) memungkinkan pengelolaan data berskala masif dengan performa tinggi dan biaya rendah.
      (5) Data Warehouse Modern: Digunakan untuk analitik terstruktur, seperti Greenplum Database atau Vertica.
   b) Karakteristik Utama Penyimpanan Big Data:
      (1) Skalabilitas: Kemampuan untuk menambah kapasitas dari terabyte ke petabyte dengan mudah.
      (2) Fleksibilitas: Mampu menyimpan berbagai jenis data: terstruktur, semi-terstruktur (JSON/XML), dan tidak terstruktur (video/gambar).
      (3) Aksesibilitas: Menyediakan antarmuka andal untuk kueri dan pemrosesan real-time.

(4) Efisiensi Biaya: Sering kali menggunakan hard disk drive (HDD) atau sistem hybrid untuk mengoptimalkan biaya penyimpanan data dalam volume besar.

V. Alur System Cerdas (AI Pipeline)

AI Pipeline adalah serangkaian proses otomatis yang terstruktur untuk mengelola seluruh siklus hidup pengembangan kecerdasan buatan, mulai dari pengumpulan data mentah, prapemrosesan, pelatihan model, evaluasi, hingga penerapan (deployment) secara real-time. Ini adalah alur kerja sistematis yang mengubah data menjadi prediksi atau tindakan, sering digunakan dalam Machine Learning Ops (MLOps) untuk efisiensi dan konsistensi.

A. Tahapan Umum dalam AI Pipeline:

1. Pengambilan & Pemrosesan Data: Mengumpulkan data mentah dari berbagai sumber, membersihkannya, dan mengubahnya menjadi format yang siap digunakan.
2. Rekayasa Fitur (Feature Engineering): Menyeleksi atau membuat variabel fitur yang relevan untuk meningkatkan akurasi model.
3. Pelatihan Model (Model Training): Algoritma AI dilatih menggunakan data yang telah diproses.
4. Evaluasi & Validasi: Menguji kinerja model untuk memastikan akurasi dan mencegah overfitting.
5. Penerapan (Deployment): Mengintegrasikan model yang sudah dilatih ke dalam aplikasi produksi.
6. Pemantauan & Pelatihan Ulang (Monitoring & Retraining): Memantau kinerja model secara real-time dan melatih ulang jika akurasi menurun.

B. Mengapa AI Pipeline Penting?

1. Otomatisasi: Mengurangi tugas manual yang rumit dan berulang.
2. Reproduksibilitas: Memastikan hasil yang konsisten setiap kali model diperbarui.
3. Skalabilitas: Memungkinkan penanganan volume data yang besar dan kompleks secara efisien.
4. Kecepatan: Mempercepat siklus pengembangan dari ide hingga ke produksi.

C. Pipeline yang diimplementasikan di Chatbot Intelijen untuk konsultasi tentang DTSEN

1. Offline Processing (Saat Upload Dokumen)
   a) Document Loading - Baca PDF / DOCX / TXT
   b) Text Extraction - Ambil teks dari file
   c) Chunking Strategy - Potong teks menjadi bagian 500 kata
   d) Embedding Creation - Ubah teks jadi vector 384D dengan SBERT

e) Vector Storage - Simpan di FAISS index

2. Online Inference (Saat User Bertanya)
   a) Query Processing : user input diterima
   b) Query Embedding : konversi query ke vector
   c) Semantic Search : cari similarity di FAISS
   d) Context Retrieval : ambil 3 dokumen paling relevan
   e) Response Assembly : format dengan template RAG
   f) UI Delivery : tampilkan di streamlit chat

VI. Implementasi AI Modern

Beberapa teknik AI modern yang coba diimplementasikan di chatbot yang saya buat diatas adalah sebagai berikut:

A. RAG (Retrieval-Augmented Generation)

1. Pattern terbaru untuk QA berbasis dokumen
2. Grounding response pada dokumen yang di upload
3. Hindari hallucination dengan konteks nyata

B. Transformer - based Embeddings

1. Sentence-BERT multilingual (paraphrase-multilingual-MiniLM-L12-v2)
2. Semantic search bukan keyword matching
3. Cross-lingual capabilities untuk bahasa indonesia

C. Vector Database & Similiarity Search

1. FAISS (Facebook AI Similarity Search) untuk pencarian cepat
2. Dense vector representations 384 dimensi
3. Cosine similarity untuk relevansi

D. Modern NLP Pipeline

1. Chunking strategy dengan overlap untuk konteks
2. Hybrid retrieval dengan semantic + metadata
3. Context-aware generation dengan template prompting

E. MLOps Practices

1. Model caching di session state
2. Vector index persistence (save/load FAISS)
3. Realtime monitoring pipeline status

Script lengkap tersedia di repository :
https://github.com/Kurnia-Wahidin/uas_smart_data_ai_system