Pembangunan *Smart Chatbot* berbasis *Natural Language Processing* untuk Layanan Penyediaan Data

Studi kasus: BPS Provinsi Sumatera Utara

Rio Manuppak Siahaan (222112324, 4SD3)

Dosen Pembimbing: Budi Yuniarto, S.S.T., M.Si.

***Ringkasan*— Badan Pusat Statistik Provinsi Sumatera Utara (BPS Sumut) bertujuan untuk meningkatkan aspek profesionalisme dengan meningkatkan pelayanan penyediaan data oleh Pelayanan Statistik Terpadu. Penggunaan *chatbot* berbasis *Large Language Model* menjadi salah satu cara untuk meningkatkan pelayanan yang interaktif. Model yang dipilih adalah Gemini 2.0 Flash Lite dan Llama 3.1 8B Instant. Evaluasi dari model terdiri atas *error rate*, ROUGE-1, ROUGE-L, *precision*, *recall*, dan *F1-score*. Dari penelitian ini nilai *error rate* pada model Gemini dan Llama berturut-turut adalah 0.0182 dan 0.0213. Nilai ROUGE-1 dan ROUGE-L Gemini berturut-turut adalah 0.5202 dan 0.4018. Nilai ROUGE-1 dan ROUGE-L Llama berturut-turut adalah 0.6598 dan 0.5333. Nilai *precision*, *recall*, *F1-score* Gemini berturut-turut adalah 0.7465, 0.9818, dan 0.8482 dan nilai untuk model Llama berturut-turut adalah 0.8519, 0.9787, dan 0.8482. Evaluasi tersebut menunjukkan bahwa Llama dapat memahami konteks lebih baik tetapi Gemini dapat mencari dokumen yang relevan daripada Llama. Dari nilai *F1 score*, kedua model memiliki kemampuan yang sama dalam menjawab secara relevan dengan data tabel statis BPS Sumut.**

***Kata Kunci*— *Chatbot*, LLM, aplikasi, BPS, PST**

I. Latar Belakang

Badan Pusat Statistik (BPS) merupakan lembaga non-kementerian yang tersebar di seluru provinsi di Indonesia berdasar Undang-undang Republik Indonesia Nomor 16 tahun 1997 memuat salah satu peranan BPS, yaitu menyediaka kebutuhan data bagi pemerintahan dan masyarakat. BPS mengumpulkan data yang terbagi menjadi data primer yang diperoleh dari sensus, survei, dan berbagai kegiatan statistik yang dilakukan oleh BPS dan data sekunder yang diperoleh dari departemen atau lembaga pemerintahan lainnya. Adapun beberapa produk yang dihasilkan oleh BPS di antaranya tabel dinamis, berita resmi statistik (BRS), publikasi, infografis, glosarium, dan masih banyak lagi. Data tersebut digunakan dalam berbagai sektor, seperti sektor ekonomi, sektor kependudukan, dan sektor pertanian.

Sebagai penyedia data yang dapat diandalkan, BPS harus mengikuti perkembangan Ilmu dan Pengetahuan dan Teknologi (IPTEK)[1]. BPS juga menyediakan sarana bagi pengguna data dalam mendapatkan data primer miliki BPS dengan menyediakan pelayanan statistik terpadu (PST). Adapun sistem pelayanan terpadu memiliki tujuan untuk: 1) melindungi dan memberikan kepastian hukum pengguna data; 2) melayani masyarakat; 3) memberikan pelayanan yang cepat; dan 4) melayani masyarakat lebih luas [2].

BPS Sumut juga mendukung tujuan dari PST dengan membuka berbagai layanan statistik secara daring yang dapat dicapai oleh masyarakat luas. Dengan banyaknya produk yang dihasilkan oleh BPS Sumut membuat pencarian data secara spesifik cenderung sulit. Pengguna data harus mengetahui subjek dan sub-subjek data yang ingin dicari dalam klasifikasi data BPS Sumut. Di lain sisi, klasifikasi data tersebut menjamin relevansi dan spesifikasi data. Melalui analisis hasil survei kebutuhan data oleh BPS tahun 2023, jenis layanan di PST yang sering diakses oleh pengguna data adalah layanan statistik dengan nilai sebesar 33,61% [32]. Berdasarkan wawancara secara daring dengan narasumber dari BPS Sumut, penggunaan *chatbot* dapat memenuhi tujuan dalam memenuhi aspek profesionalisme dengan layanan 24 jam.

Pelayanan data selama ini dilakukan secara tatap muka atau melalui *chat* dari dari web yang dilakukan pada jam kantor sekitar pukul 08.00 WIB sampai 15.30 WIB. Dengan implementasi *Chatbot*, pelayanan permintaan data dapat dilakukan selama 24 jam. Berdasarkan survei milik Brandtzaeg [38], sebesar 68% responden menggunakan *Chatbot* dalam membantu meningkatkan produktivitas dari sisi kemudahan, kecepatan, dan kenyamanan dalam memperoleh bantuan dan informasi. Integrasi dengan LLM membuat pengguna data dapat memperoleh data menggunakan bahasa sehari-hari. Oleh karena itu, aplikasi *Chatbot* harus mampu mengelola kata dari berbagai bahasa sehingga dapat dipakai oleh pengguna data dengan skala luas. Dengan aplikasi *Chatbot* ini, pengguna data dapat meminta data secara 24 jam dan dimana saja.

Aplikasi *Chatbot* dapat memenuhi pelayanan permintaan data statistik. Oleh karena itu, *Chatbot* memiliki data berupa tabel dinamis dan berbagai fail PDF tentang *official statistics* dan BPS. Dengan adanya aplikasi ini, pegawai PST juga dapat mengawasi transaksi permintaan data melalui halaman administrator sehingga pihak BPS Sumut dapat mengetahui tabel yang sering diminta oleh pengguna data.

Implementasi *natural language processing* (NLP) ditujukan untuk meningkatkan interaktivitas *Chatbot*. NLP adalah salah satu cabang kecerdasan buatan yang berfokus pada interaksi manusia dan komputer menggunakan bahasa alami. Adapun beberapa penerapan NLP dalam berbagai bidang di antaranya *natural language user interface*, *automated text summarization*, *information extraction*, *text mining*, dan

document retrieval [3]. *Text generation*, *part-of-speech tagging* dan *parsing* adalah beberapa komponen NLP [4].

Large language model (LLM) adalah salah satu pengembangan dari teknologi NLP berbasis kecerdasan buatan yang telah digunakan secara global [5]. LLM adalah model *deep learning* yang dilatih dengan korpus yang besar sehingga dapat memahami dan menghasilkan bahasa yang alami [6]. LLM memiliki arsitektur neural network yang disebut Transformers yang mendukung dalam memahami bahasa yang kompleks dan menghasilkan bahasa yang alami. Untuk itu, LLM dilatih dengan dataset besar dengan milyaran kata dari berbagai sumber seperti situs web dan buku [7].

II. Tujuan Penelitian

Tujuan umum dari penelitian ini membangun aplikasi *smart chatbot* berbasis NLP untuk pelayanan penyediaan data. Adapun tujuan khusus dari penelitian ini adalah sebagai berikut.

1. Menerapkan perhitungan similaritas kata untuk pencarian semantik data tabel dinamis BPS Provinsi Sumatera Utara.
2. Mengimplementasikan dan mengevaluasi model LLM pada aplikasi *Chatbot* untuk menyediakan data dari pertanyaan kompleks pengguna terkait data tabel dinamis BPS Provinsi Sumatera Utara.
3. Membangun dan mengevaluasi sistem aplikasi *Chatbot* dan *dashboard* pengawasan penggunaan *Chatbot* dan modifikasi *knowledge base*.

III. Penelitian Terkait

Perkembangan IPTEK telah membawa jalan baru dengan kecerdasan buatan. Kecerdasan buatan ini kemudian diimplementasikan dalam aplikasi *question answering chatbot*. Integrasi *chatbot* berbasis AI dapat digunakan untuk berbagai keperluan [10-14]. Menurut kamus dari Lexico Dictionaries, *chatbot* adalah sebuah program komputer yang mensimulasikan percakapan manusia, terkhusus dari internet [15]. Bot menggunakan natural language processing (NLP) dan analisis sentimen untuk berkomunikasi dalam bahasa manusia dengan teks atau percakapan dengan manusia atau dengan bot lainnya [16]. NLP adalah cabang dari sains data dan kecerdasan buatan yang fokus untuk memahami bahasa manusia. Oleh karena itu, NLP dikombinasikan dengan pendekatan linguistik, ilmu komputer, dan statistika dalam penyelesaian tugas yang berhubungan dengan bahasa, seperti klasifikasi teks, analisis sentimen, dan pembuatan bahasa [17- 19].

Saat ini, banyak sekali model LLM yang dapat digunakan untuk berbagai tugas. Salah satunya model Gemin
Saat ini, kecerdasan buatan dengan implementasi LLM telah meningkatkan kapabilitas *chatbot*. Kecerdasan buatan ini dilatih dengan dataset yang masif dan memanfaatkan teknik NLP untuk menghasilkan respons dalam bahasa manusia. Contoh dari LLM adalah GPT4 milik OpenAI dan Gemini dan Google Deepmind. Model GPT4 sering digunakan dalam berbagai sistem informasi, salah satunya kesehatan, dimana model ini dapat membantu dokter untuk menjawab pertanyaan dari pasien selama PET/CT [20-22]. Model di atas merupakan kemajuan besar kecerdasan buatan yang telah diluncurkan. Adapun model lainnya, seperti LaMDA [23], GPT-NeoX[24], PaLM2 [25], dan LLaMA 2[26]. Model tersebut dapat memahami, menghasilkan, dan berinteraksi dengan bahasa manusia [27]. LLM dapat menghasilkan dan menganalisis teks, secara efektif mengelola informasi penting dan kompleks [28]. *Chatbot* memiliki dataset sebagai dasar untuk latihan dan generasi teks. Dataset tersebut memiliki struktur tidak teratur yang dapat bersumber dari buku teks, slide, transkrip kuliah, penelitian makalah, dan lainnya. Untuk bisa membaca dataset tersebut, LLM memiliki cara untuk bisa mengatasi tantangan tersebut dengan metode retrieval-augmented generation (RAG) [29]. Pendekatan RAG dapat mengekstraksi informasi yang relevan secara efisien dari dataset yang tidak terstruktur dari pertanyaan pengguna. LLM dengan metode RAG dapat menjadi jawaban pada sistem *Chatbot* yang memiliki dasar dataset. Sistem berbasis RAG dapat mengambil informasi secara dinamis dari kumpulan data dan menghasilkan respons yang disesuaikan dengan permintaan pengguna [30]. Pembangunan sistem ini menggunakan framework LangChain [31].

IV. Kerangka Pikir

Alur pemikiran dari penelitian ini seperti pada Gambar 1. Kerangka pikir pada penelitian ini berawal dari berbagai data yang dihasilkan oleh BPS Sumut membuat pengguna data kesulitan dalam mencari data secara daring. Adapun pelayanan secara luring hanya bisa dilakukan pada jam kantor. Dengan adanya *Chatbot*, pelayanan dapat dilakukan secara daring selama 24 jam. Aplikasi *Chatbot* ini mengimplementasi model bahasa besar (LLM) sehingga dapat mengelola bahasa sehari-hari. Pengguna data tidak hanya bisa mencari data secara daring, juga meminta data secara efektif dan mudah menggunakan bahasa sehari-hari.

BPS Sumut juga dapat memenuhi aspek profesionalisme SDM, yaitu peningkatan pelayanan publik dengan menyediakan layanan secara daring dan 24 jam. BPS Sumut juga dapat mengawasi permintaan data dan mengelola data transaksi tersebut untuk membuat keputusan dalam pengelolaan data ke depannya.

V. Penelitian

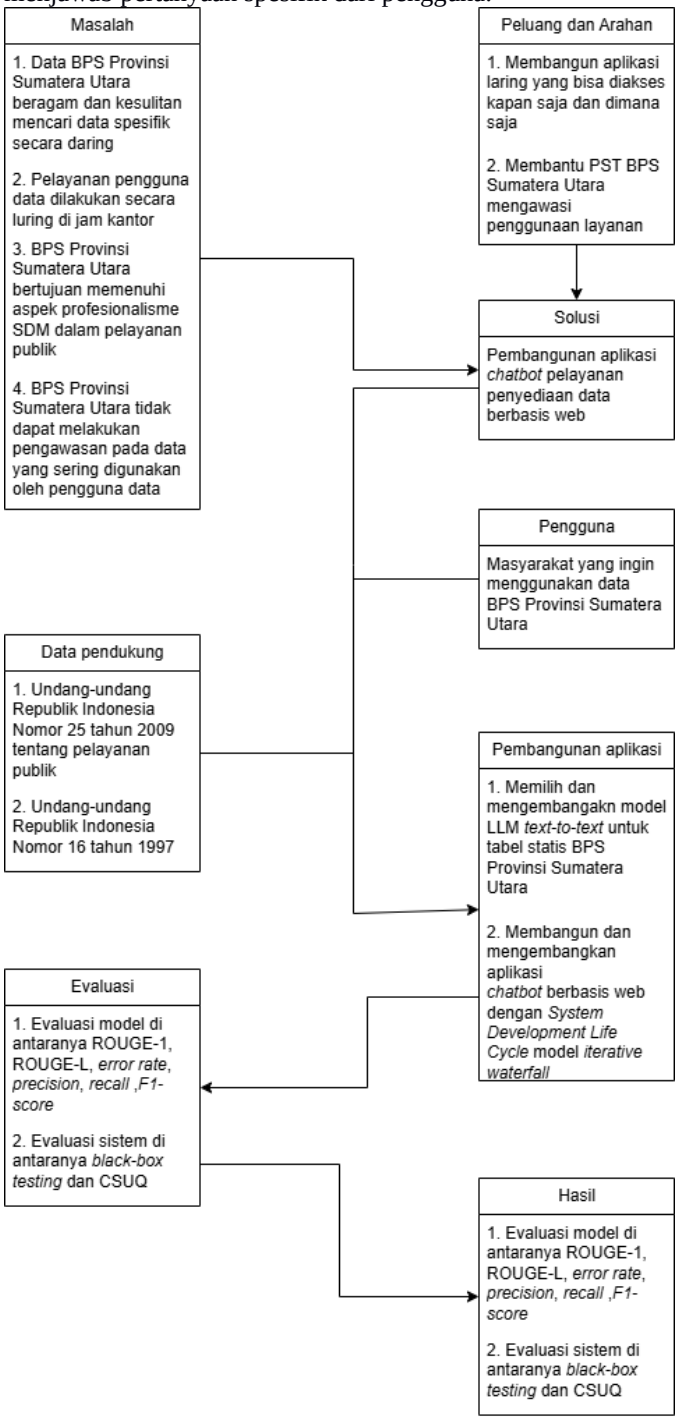
A. Ruang Lingkup Penelitian

Penelitian ini terbatas pengembangan sistem *Chatbot* berbasis NLP untuk penyediaan data tabel dinamis BPS Provinsi Sumatera Utara. Sistem *Chatbot* ini ditujukan untuk pengguna data yang ingin mengakses data tabel dinamis BPS Provinsi Sumatera Utara. Adapun sistem *dashboard* untuk administrator yang merupakan pegawai PST BPS Provinsi Sumatera Utara. Administrator (admin) dapat mengawasi penggunaan layanan dan mengelola *knowledge base* untuk model.

B. Metode Pengumpulan Data

Data yang dikumpulkan merupakan data sekunder yang berasal dari BPS Provinsi Sumatera Utara. Data yang dikumpulkan pada penelitian ini berfokus pada data tabel dinamis dari website BPS Provinsi Sumatera Utara. Pengumpulan data dilakukan melalui *application programming interface* (API) milik BPS sendiri. Data tersebut akan menjadi

knowledge base yang mempengaruhi kemampuan model menjawab pertanyaan spesifik dari pengguna.



Gambar 1. Kerangka pikir

Data yang diambil akan digunakan sebagai konten untuk model dapat merespons permintaan data. Data yang digunakan adalah data sekunder yang berasal dari web API BPS untuk layanan permintaan data. Pengumpulan data dari web API BPS menggunakan metode scrapping. Metode ini memungkinkan untuk pengambilan data secara otomatis sehingga memangkas waktu. Pengembang sistem (*developer*) harus melewati tahap register dan login terlebih dahulu. Developer yang sudah terautentikasi akan mendapatkan API Key untuk melakukan requests ke API. Respons dari API akan berupa data berformat JSON yang akan disimpan di lokal. Format JSON lebih mudah untuk diolah menjadi format data yang diinginkan, di antaranya menjadi format Excel atau format CSV. Perubahan format ini mempermudah dalam pengembangan sistem. Adapun parameter yang perlu diisi adalah parameter model yang diisi "data", parameter domain yang diisi 1200 (domain milik BPS Provinsi Sumatera Utara), parameter key yang idisi dengan API Key, parameter var yang diisi secara menaik dari angka 1 sampai 700 (diasumsikan tabel dinamis untuk domain BPS Provinsi Sumatera Utara sampai 700). Data dengan format JSON diubah menjadi data berformat CSV. Fail CSV akan di input melalui GUI, seperti streamlit terlebih dahulu. Lalu, fail teks di CSV diubah menjadi vektor oleh model *embedd* dan disimpan dalam basis data vektor.

C. Metode pengelolaan data

1. *Preprocessing*

Pada tahapan ini, fail yang dikumpulkan berupa JSON akan diubah menjadi fail CSV. Perubahan struktur fail juga mempengaruhi struktur data. Adapun data yang diambil dari struktur fail JSON adalah data tahun, data turvar, data vervar, data *datacontent*. Data turvar berisi kolom dari tabel. Data vervar berisi baris dari tabel. Data *datacontent* adalah data numerik.

2. Input fail CSV

Setelah struktur fail JSON menjadi struktur fail CSV berbentuk tabel, fail tersebut diinput dalam GUI untuk diubah menjadi bentuk vektor. Jumlah fail CSV sebanyak 477 fail yang memuat 477 tabel dinamis dari BPS Sumatera Utara. GUI yang dipakai berbentuk web dan dapat diakses oleh pegawai BPS yang bertanggung jawab dalam menambah *knowledge base* pada model. *Knowledge base* adalah dasar pengetahuan yang akan dipakai oleh LLM dalam memberikan jawaban sehingga jawaban sesuai dengan data BPS yang diinput.

3. Menambah dokumen ke *vector database*

Fail diterima akan secara iterasi dimasukkan ke dalam *vector database* dengan model *embedding* dari Gemini. Model *text-embedding-004* adalah model *embed* dengan keunggulan dari model *embed* dengan jumlah dimensi input dan dimensi output yang lebih baik. Nilai dari MTEB sebesar 66.31% dengan jumlah token input yang lebih kecil sebesar 2048. Menurut Jinhyuk Lee, dkk., model embedd ini dapat bersaing 7 kali dari model yang lebih besar dan 5 kali dari model dengan dimensi lebih banyak [8].

D. Pengembangan model

1. Model yang digunakan

Model LLM yang dipilih adalah model LLM open-source hingga berbayar. Model LLM pada penelitian ini adalah model untuk *text-generation dan text-embedding*. Model yang digunakan sebagai *text-generation* adalah dengan membandingkan dua *pre-trained* model, yaitu Gemini 2.0 Flash Lite dan Llama 3.1 8B Instant.

2. Evaluasi model

Terdapat beberapa evaluasi model yang digunakan dalam penelitian ini. Untuk menentukan model yang terbaik dalam menghasilkan teks, dilakukan tes ROUGE-1 dan ROUGE-L berdasarkan konten dari augmentasi kueri out-of-domain (OOD) [33]. Evaluasi dilakukan dengan membentuk input

prompt dan *ground truth* secara *custom*. Lalu, *prompt* yang dibentuk menjadi input bagi model untuk menjawab sesuai dengan *knowledge base*. Jawaban dari model akan direkam sebagai *generated answer*. *Ground truth* dan *generated answer* akan digunakan untuk tes ROUGE-1 dan ROUGE-L.

Evaluasi lainnya adalah *precision, recall, F1 score*. Evaluasi ini digunakan untuk menguji performa dari jawaban model berdasarkan dokumen yang disediakan (*knowledge base*). Adapun persamaan dari *precision, recall, F1 score*.

$$Precision = \frac{TP}{FP+TP} \quad \text{.....(1)}$$

$$Recall = \frac{TP}{FN+TP} \quad \text{.....(2)}$$

$$F1 - score = 2 \times \frac{precision \times recall}{precision+recall} \quad \text{.....(3)}$$

Dengan keterangan sebagai berikut. *True positive* (TP) adalah kondisi jawaban model yang sama dengan *ground truth*. *False positive* (FP) adalah kondisi jawaban model tidak sesuai dengan jawaban yang *ground truth*. *False negative* (FN) adalah kondisi model menjawab tidak ada yang sebenarnya ada pada *ground truth*.

Evaluasi jawaban model berdasarkan *knowledge base* juga dilakukan menghitung *error rate*. Semakin kecil nilai dari *error rate* menunjukkan risiko kesalahan jawaban model kecil. Adapun persamaan dari *error rate* adalah sebagai berikut.

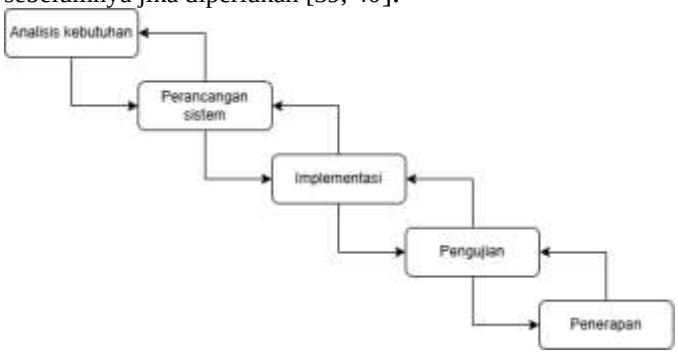
$$error\ rate = \frac{TN}{TN+TP} \quad \text{.....(4)}$$

E. Pembangunan sistem

1. Metode pembangunan sistem

Metode pembangunan sistem aplikasi *Chatbot* yang digunakan pada penelitian ini adalah *Software Development Life Cycle* (SDLC) dengan model *iterative waterfall*. Metode ini dianggap sebagai pengembangan sistem dengan persyaratan kebutuhan yang stabil dan meminimalkan perencanaan yang berlebihan karena perencanaan sudah dilakukan di awal [39].

Adapun tahapan dari SDLC model *iterative waterfall*, antara lain analisis kebutuhan, perancangan sistem, implementasi, pengujian dan penerapan. Tahapan model dapat dilakukan secara berurutan dan dapat kembali ke langkah sebelumnya jika diperlukan [39, 40].



Gambar 2. Tahapan SDLC model *iterative waterfall*

Tahapan analisis kebutuhan dilakukan dengan melakukan wawancara dengan perwakilan dari PST BPS Provinsi Sumatera Utara yang merupakan *subject matter* dari penelitian ini. Wawancara tersebut bertujuan untuk mengidentifikasi kebutuhan dari sistem yang harus terpenuhi. Metode analisis kebutuhan yang lain adalah melakukan analisis fitur dari sistem *chatbot* yang sudah ada di BPS Provinsi Sumatera Utara.

Tahapan perancangan sistem adalah mendaftarkan seluruh kebutuhan yang diperlukan dalam sistem dan menyusun rancangan proses bisnis, rancangan arsitektur, diagram *use case*, dan rancangan antarmuka sistem usulan. Rancangan antarmuka akan berupa prototipe yang akan menjadi landasan pengembangan sistem ke depannya. Prototipe tersebut mampu memberikan gambaran tampilan dan fitur yang akan digunakan dalam sistem.

Tahapan implementasi merupakan tahapan pembangunan yang berdasarkan rancangan sistem sebelumnya. Sistem aplikasi yang dibentuk adalah sistem *chatbot* berbasis LLM untuk layanan penyediaan data. Sistem akan menggunakan dua bahasa pemrograman yang dibagi menjadi bahasa python untuk pengembangan model dan FastAPI sebagai penghubung dan bahasa JavaScript dengan *framework* ReactJS, yaitu NextJS sebagai tampilan antarmuka.

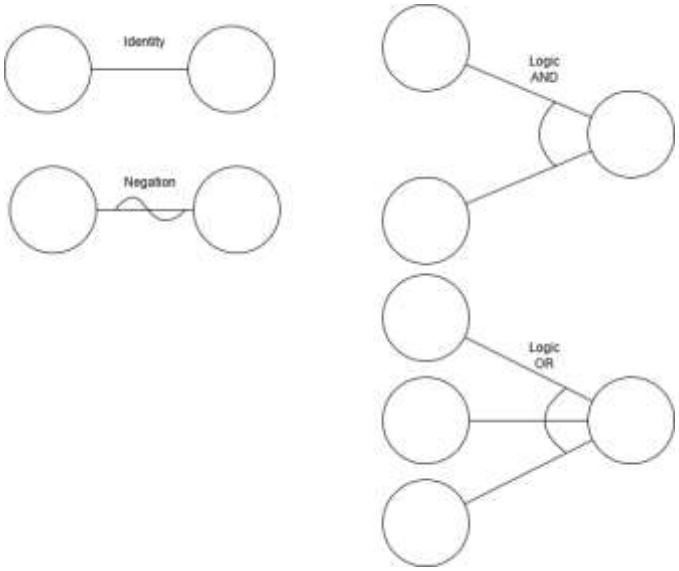
Setelah tahapan implementasi, pengujian dilakukan untuk melihat apakah sistem dapat berjalan dan memenuhi kriteria sesuai dengan rancangan sebelumnya. Pengujian akan dilakukan dengan dua metode, yaitu *black-box testing* dan *usability testing* dengan CSUQ.

2. Evaluasi sistem

Terdapat metode evaluasi yang diterapkan pada pengujian sistem. Evaluasi tersebut di antaranya pengujian *black-box testing* dan disusul dengan pengujian menggunakan sistem *Computer System Usability Questionnaire* (CSUQ).

Pengujian *black-box* ditujukan untuk menguji kesesuaian fungsionalitas dari aplikasi yang dibangun. Pengujian dilakukan oleh penguji dengan satu atau lebih yang akan penerapan skenario pengujian terhadap fitur aplikasi. Pengujian *black-box*, penguji tidak perlu mengetahui *source code* atau sistem kerja belakang layar atau internal [40, 41]. Teknik yang digunakan untuk membentuk kasus uji pada pengujian ini berupa *cause-effect graph*. Royce [40] menjelaskan bahwa *cause-effect graph* merupakan teknik *black-box testing* dimana pengujian dimulai dengan pembuatan grafik dan menetapkan hubungan sebab akibat. *Cause-effect graph* terdiri dari empat simbol dasar sebagai berikut.

Metode berikutnya, yaitu CSUQ bertujuan untuk mengetahui kepuasan pengguna akhir menggunakan aplikasi. CSUQ dapat mengukur kepuasan pengguna secara multidimensi [41, 42]. CSUQ juga tidak memerlukan biaya lisensi sama sekali [43]. Kuesioner CSUQ ini terdiri dari 16 item pernyataan yang termuat di dalam empat dimensi, yaitu *Overall*, *System Usefulness* (SysUse), *Information Quality* (InfoQual), *Interface Quality* (IntQual). Hasil dari CSUQ berupa numerik dari 1 sampai 7 dimana nilai 1 menunjukkan responden sangat setuju dan nilai 7 menunjukkan responden sangat tidak setuju [43]. Skor tersebut dapat menjadi acuan peningkatan kualitas dan perbaikan sistem aplikasi dari sisi kegunaan, informasi, maupun antarmuka.



Gambar 3. Simbol dasar *cause-effect graph*

VI. HASIL DAN PEMBAHASAN

A. Analisis masalah

BPS Provinsi Sumatera Utara (BPS Sumut) memiliki variasi data yang besar dimulai dari tabel dinamis, hasil sensus dan survei, berita resmi statistik (BRS), publikasi, infografis dan sebagainya. Pada penelitian ini berfokus pada data tabel dinamis milik BPS Provinsi Sumatera Utara yang memiliki berbagai subjek dan sub-subjek untuk mengklasifikasikan data. Hal tersebut baik untuk menempatkan data sesuai dengan kategori masing-masing. Di lain sisi, hal tersebut dapat mempersulit pengguna data yang ingin mengakses data melalui web BPS Sumut. Pengguna data harus datang secara luring ke dalam bagian Pelayanan Statsitik Terpadu (PST) BPS Sumut untuk mendapatkan data yang dicari. Adapun BPS Provinsi Sumatera Utara berupaya memenuhi aspek profesionalisme SDM dalam pelayanan publik. Salah satu caranya adalah meningkatkan jam pelayanan menjadi 24 jam.

Hal tersebut dapat dipermudah dengan adanya aplikasi *Chatbot* berbasis NLP pada web sehingga dapat diakses dengan mudah selama 24 jam. Penggunaan NLP pada *Chatbot* ditujukan untuk pengguna data dapat berinteraksi menggunakan bahasa sehari-hari karena *Chatbot* dapat mengelola berbagai bahasa manusia. Pihak PST pun dapat mengawasi transaksi antara *Chatbot* dan pengguna data. Pihak PST pada aplikasi ini disebut *admin* daapt mengetahui data yang sering diminta oleh pengguna data dan jumlah penggunaan aplikasi mingguan sampai bulanan. Admin juga dapat menambahkan, menghapus, dan mengunduh fail untuk keperluan model.

B. Analisis kebutuhan

Kebutuhan fungsional yang dipenuhi sistem berdasarkan analisis masalah sebelumnya dan tambahan fungsional untuk meningkatkan kepuasan pengguna adalah kemampuan menganalisis data yang diminta oleh pengguna data, kemampuan personifikasi model sesuai dengan pengguna aplikasi, pencarian semantik yang akurat berdasarkan similaritas data, kemampuan menjawab pertanyaan secara spesifik terkait tahun dan wilayah berdasarkan *knowledge base*, pilihan tampilan antarmuka antara mode terbang dan mode gelap, kemampuan segmentasi obrolan, pemilihan model LLM yang akan digunakan untuk menjawab pertanyaan, dan kemampuan mengelola *knowledge base* model, dan menyapa pengguna terlebih dahulu berdasarkan nama pengguna.

C. Desain sistem usulan

Sesuai dengan analisis kebutuhan sebelumnya, berikut perancangan proses bisnis sistem. Pengguna data wajib terautentikasi terlebih dahulu sebelum menggunakan *Chatbot* melalui sistem registrasi dan *login*. Setelah terautentikasi, pengguna diarahkan ke halaman *chat* untuk menggunakan layanan *Chatbot* INDA. Input dari pengguna akan dikirim ke model NLP menggunakan FastAPI. Hasil generasi teks oleh model akan dikirim kembali oleh FastAPI ke UI. Respon yang diterima akan menjadi respon *error* jika terkait dengan penggunaan melebihi *rate limit* dan sebagainya. Pengguna dapat memilih model LLM dan mengatur tema terang atau gelap untuk tampilan percakapan.

D. Pengumpulan data

Data dikumpulkan dengan melakukan *request* secara iterasi ke URL RestAPI dari Web API BPS (https://webapi.bps.go.id). Data yang diambil adalah data tabel dinamis BPS Provinsi Sumatera Utara (BPS Sumut). Parameter yang diperlukan untuk melakukan *request* adalah *model* dengan input *data*, *lang* dengan inputan *id*, *domain* dengan input 1200 sebagai kode wilayah BPS Sumut, *var* adalah nomor tabel dinamis, input *turvar, vervar, th*, dan *turth* bersifat opsional, dan *key* dengan input API *key* pengguna. *Var* diisi dengan input 1 sampai 700 secara iterasi. Respon dari *request* adalah fail JSON. Fail JSON akan disaring mengambil label dari *subject, var, turvar, vervar, tahun*, dan nilai dari *datacontent*. Hasil saring memiliki format CSV dan akan digunakan sebagai *knowledge base* bagi model.

TABEL I
EVALUASI MODEL LLM

Jenis Evaluasi	Detail	Gemini 2.0 Flash Lite	Llama 3.1 8B Instant
(1)	(2)	(3)	(4)
Robustness	*Error rate*	0.0182	0.0213
ROUGE	ROUGE-1	0.5202	0.6598
	ROUGE-L	0.4018	0.5333
Precision		0.7465	**0.8519**
Recall		**0.9818**	0.9787
F1 score		0.8482	0.8482

TABEL II
RATE LIMITS

Model	**Rate limits**	***Free of charge***	***Pas as you go / Developer Tier***
Gemini 2.0 Flash Lite	RPM	30	4000
	TPM	1000000	4000000
	RPD	1500	-
Llama 3.1 8B Instant	RPM	30	1000
	TPM	6000	250000
	RPD	14400	500000

Terdapat total 487 fail CSV hasil konversi dari JSON hasil respon API Web BPS.

E. Implementasi basis data vektor

Seperti yang dijelaskan sebelumnya, konten yang dibentuk akan diubah menjadi vektor dan disimpan dalam basis data vektor. Basis data vektor yang digunakan dalam penelitian ini adalah Qdrant dengan nama koleksi adalah *inda_collection*. Konversi menjadi vektor menggunakan *text-embedding-004* dari Gemini dengan ukuran dimensi output 768.

F. Implementasi dan evaluasi model LLM

Model LLM pada penelitian ini adalah Gemini 2.0 Flash Lite dan Llama 3.1 8b Instant yang merupakan model *closed source*. Kedua model akan diuji dan dibandingkan untuk mengetahui kemampuan model dalam memberi jawaban sesuai *knowledge base*. Kedua model tersebut dapat dipakai oleh pengguna data pada halaman percakapan. Model akan dipanggil dengan *interface* LangChain. Adapun *template* untuk *prompt* untuk membantu model memberikan jawaban sesuai dengan input pengguna dan konteks yang diberikan, serta dapat menjawab sesuai dengan personifikasi pengguna. Model akan mencari 50 dokumen berdasarkan tingkat similaritas tertinggi dengan input pengguna.

Hasil generasi model LLM dari 50 *top retrieval document* dievaluasi dengan beberapa pendekatan, yaitu *precision*, *recall*, *F1 score*, *error rate*, dan ROUGE-1, serta ROUGE-L yang dapat dilihat pada Tabel II. Model Gemini memiliki *error rate* lebih kecil dari model Llama mencerminkan kemampuan dalam mengidentifikasi jawaban relevan. Pada uji ROUGE-1 dan ROUGE-L, model Llama menujukkan hasil yang lebih baik dari model Gemini mengindikasikan model Llama dapat memahami konteks dengan lebih baik. Nilai *precision* model Llama lebih tinggi dari Gemini mencerminkan kemampuan Llama memberikan jawaban yang sesuai dengan *knowledge base*. Nilai *recall* model Gemini lebih tinggi dari Llama mencerminkan kemampuan Gemini mencari dokumen yang relevan dengan input pengguna. Nilai *F1 score* pada kedua model sama mencerminkan kemampuan kedua model sama baik dalam menangani data tabel dinamis BPS Sumut.

Model *closed source* memiliki penggunaan yang terbatas atau *rate limit*. *Rate limit* terdiri *requests per minute* (RPM), *requests per day* (RPD), *tokens per minute* (TPM). Ketika penggunaan dilakukan melebihi *rate limit* akan mengembalikan respon *error*. *Rate limit* dari model Gemini 2.0 Flash Lite dan Llama 3.1 8B Instant dapat dilihat dari tabel III. Model Gemini dapat ditingkatkan dengan menyusun *plan* dan *set up billing*. Biaya *pay-as-you-go* untuk model Gemini dibagi untuk input dan output tiap satu juta token dalam USD. Biaya input adalah $0.075 untuk teks, gambar, video, dan audio, dan biaya output adalah $0.30. Biaya *developer tier* untuk model Llama 3.1 dari Groq adalah $0.05 tiap 20 juta input token dan $0.08 tiap 12.5 juta token.

G. Pengembangan sistem

Aplikasi *web Chatbot* diberi nama INDA (*Intelligent Data Assistant*) dibangun sesuai dengan fitur yang dimiliki oleh *Electronic Data Assistant* (EDA). EDA adalah aplikasi *Chatbot* milik BPS Sumut yang memberikan *link* relevan dengan permintaan pengguna. Pengguna melakukan interaksi dengan EDA menggunakan kode yang disediakan oleh EDA sehingga tidak interaktif dan responsif terhadap pengguna. Penyediaan *link* yang diberikan terbatas dan tidak bisa memberikan jawaban langsung data bagi pengguna. Dengan penggunaan NLP, INDA dapat memberikan jawaban yang relevan secara langsung kepada pengguna serta memberikan jawaban yang interaktif dan menggunakan bahasa sehari-hari bagi.

Tampilan pertama dari web adalah *landing page* yang menyajikan pengertian aplikasi INDA (*Intelligent Data Assistant*). Web terdiri dari dua tampilan sesuai dengan dua *actor* sesuai autentikasi. Tampilan pengguna (*user*) memiliki tampilan utama, yaitu tampilan percakapan atau *chat*. Tampilan administrator yang ditujukan untuk pegawai PST BPS Sumut memiliki tampilan utama, yaitu *dashboard* untuk melihat grafik penggunaan layanan dan *rate limit* tiap model.

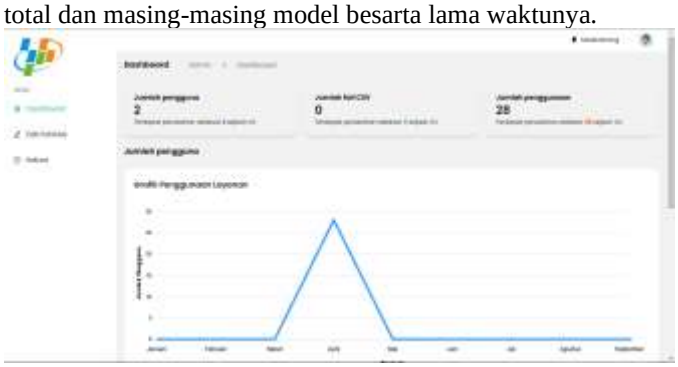
Tahap awal pengembangan sistem mencakup halaman yang responsif, fitur tema terang gelap, dan menghubungkan model LLM dengan aplikasi bagi halaman percakapan dan admin. Fitur yang dikembangkan selanjutnya adalah menambah percakapan, mengedit judul percakapan, dan menghapus percakapan. Pengembangan selanjutnya akan berfokus pada halaman administrator dengan menambah grafik penggunaan layanan dan menghitung penggunaan layanan. Fitur yang dikembangkan selanjutnya ialah fitur mengedit *knowledge base* bagi model.

H. Evaluasi sistem

Evaluasi sistem pada penelitian ini adalah *black-box testing* dan *usability testing*. Pada tahap ini, *black-box testing* menangkap hasil bahwa 20 dari 22 fitur yang berhasil memenuhi output yang diharapkan. Adapun fitur penting dari sisi pengguna adalah sistem percakapan berbasis LLM. Model dapat dipilih oleh pengguna sehingga menambah interaktivitas dengan pengguna.

Gambar 4. Tampilan halaman percakapan mode terang

Di lain sisi, fitur penting dari sisi administrator adalah *dashboard* yang dapat mengawasi penggunaan layanan secara total dan masing-masing model besarta lama waktunya.



Gambar 5. Tampilan *dashboard* mode terang

VII. PENUTUP

Berikut kesimpulan sementara yang didapatkan dari hasil penelitian di atas.

1) Model dapat menjawab pertanyaan berdasarkan *knowledge base* yang didapatkan dari API BPS. Respon dari API BPS berupa JSON diekstrak dan disederhanakan dalam format CSV lalu diubah dalam bentuk kalimat untuk mempermudah dalam model mendapatkan dokumen yang relevan dengan permintaan pengguna. Pemilihan model *embed* adalah model dari Gemini, yaitu *text-embedding-004*. Adapun evaluasi *robustness* berupa *error rate* yang didapatkan dari proses RAG yang membantu model menjawab berdasarkan *knowledge base* pada model Gemini 2.0 dan Llama 3.1 secara berturut-turut adalah 0.0182 dan 0.0213. Model Gemini memiliki nilai *error rate* lebih kecil menggambarkan model dapat mencari dan menjawab berdasarkan dokumen yang menjadi *kwowledge base* yang relevan pada permintaan pengguna.
2) Implementasi model pada tampilan UI pengguna menggunakan FastAPI. Input dari pengguna dikirim melalui FastAPI dan hasil generasi teks dari model dikembalikan ke UI pengguna. Adapun evaluasi model pada UI berupa ROUGE-1, ROUGE-L, *precision*, *recall*, dan *F1-score*. Pada Tabel II, evaluasi ROUGE didominasi oleh model Llama 3.1 yang menggambarkan model dapat memahami konteks lebih baik. Nilai *precision* model Llama lebih tinggi dari Gemini mencerminkan kemampuan Llama memberikan jawaban yang sesuai dengan *knowledge base*. Nilai *recall* model Gemini lebih tinggi dari Llama mencerminkan kemampuan Gemini mencari dokumen yang relevan dengan input pengguna. Nilai *F1 score* pada kedua model sama mencerminkan kemampuan kedua model sama baik dalam menangani data tabel dinamis BPS Sumut.
3) Evaluasi sistem *Chatbot* melibatkan dua jenis, yaitu *black-box testing* dan *usability testing*. Uji *black-box* mengindikasikan bahwa sebagian besar dari fitur yang direncanakan sudah berhasil memenuhi output yang diharapkan.

Rencana pekerjaan susulan berupa peningkatan UI pada bagian animasi dan visual web, dan peningkatan UX pada halaman percakapan dan halaman pada admin. Melanjutkan fitur yang belum selesai pada sistem dan menambahkan fitur memberikan *rating* dan komentar pada aplikasi. Uji CSUQ menjadi penutup untuk menguji sistem *Chatbot* dan menerima saran dan komentar dari penguji. Adapun perubahan model dan tahap pengembangan sistem ditujukan untuk mempermudah pengembangan sistem dan menyederhanakan tujuan dari proposal.

DAFTAR PUSTAKA

[1] "Undang-undang (uu) nomor 16 tahun 1997 tentang statistik," Republik Indonesia, 5 1997. [Online]. Available: https://peraturan.bpk.go.id/Download/34497/UU%20Nomor%2016%20Tahun%201997.pdf

[2] "Undang-undang (uu) nomor 25 tahun 2009," Republik Indonesia. [Online]. Available: https://eppid.mahkamahagung.go.id/fails/shares/uu%2025%20tahun%2209-pelayanan%20publik.pdf

[3] Russel, S. J. dkk., Artificial intelligence a modern approach. London,2010.

[4] Pustejovsky, J., A. Stubbs, Natural Language Annotation for Machine

[5] Learning: A guide to corpus-building for applications. "O'Reilly Media, Inc.," 2012.

[6] Ozdemir S., Quick Start Guide to Large Language Models: Strategic and Best Practices For Using ChatGPT and Other LLMs. Addison-Wesley Professional,2023.

[7] Peng, R. dkk., Embedding-based retrieval with llm for effective agriculture information extracting from unstructured data, 2023. [Online]. Available: arXiv preprint arXiv:2308.03107.

[8] Y. Chang dkk., A survey on evaluation of large language models, 2023. [Online]. Available: arXiv preprint arXiv:2307.03109.

[9] L. Jinhyuk, dkk.,Gecko: Versatile Text Embeddings Distilled from Large Language Models,2024. [Online]. Available: arXiv:2403.20327v1 [cs.CL].

[10] G. Siemens,Connectivism: a learning theory for the digital age. International Journal of Instructional Technology and Distance Learning, 2005, pp.3-10.

[11] X. Deng, Z. Yu, A meta-analysis and systematic review of the effect of *Chatbot* technology use in sustainable education.Sustainability, 2023,pp.2940.

[12] L. K. Fryer, D. Coniam, dkk.,Bots for language learning now: current and future directions.Lang Learn Technol,2020,pp.8-22.

[13] J. Jeon, S. Lee, dkk.,Beyond ChatGPT: a conceptual framework and systematic review of speech-recognition *Chatbot*s for language learning. Comput Educ,2023.

[14] L. Kohnke, dkk.,ChatGPT for language teaching and learning.RELC J.,2023,pp.537-550.

[15] H. Xie, dkk.,Trends and development in technology-enhanced adaptive/personalized learning: a systematic review of journal publications from 2007 to 2017.Comput Educ,2019.

[16] Lexico Dictionaries.(2019).*Chatbot* | definition of *Chatbot* in english by Lexico Dictionaries. [Online]. Available: https://www.lexico.com/en/definition/*Chatbot*

[17] Khanna, A., dkk., A study of today's A.I. through *Chatbot*s and rediscovery of machine intelligence.International Journal of U- and e-Service, Science and Technology,2015,pp.277-284.
[18] Nadkarni, dkk.,Natural language processing: an introduction.J Am Med Inform Assoc,2011,pp.544-551.
[19] Velupillai S., dkk.,Using clinical natural language processing for health outcomes research: overview and actionable suggestions for future advances.J Biomed Inform,2018,pp.11-19.
[20] Voytovich L., dkk.,Natural language processing:practical applications in medicine and investigation of contextual autocomplete.Acta Neurochir Suppl,2022,pp.207-214.
[21] Rogasch JMM., dkk.,ChatGPT: can you prepare my patients for [(18)F] FDG PET/CT and explain my reports? J Nucl Med,2023,pp.1876-1879.
[22] Wang H., dkk.,Performance and exploration of ChatGPT in medical examination,Int J Med Inform,2023.
[23] Funk MA., dkk.,Potential of ChatGPT and GPT-4 for data mining of freetext CT reports on lung cancer.Radiology,2023.
[24] Thoppilan, R., dkk., Lamda: Language models for dialog applications. ,2022.[Online]. Available: arXiv:2201.08239.
[25] Black, S., dkk.,Gpt-neox-20b: an open-source autoregressive language model,2022. [Online]. Available: arXiv:2204.06745.
[26] Anil, R., dkk.,Palm 2 technical report,2023. [Online]. Available: arXiv:2305.10403.
[27] Touvron, H., dkk.,Llama 2: Open foundation and fine-tuned chat models.2023. [Online]. Available: arXiv:2307.09288.
[28] Jawahar, G., dkk.,What does bert learn about the structure of language? In: ACL 2019-57th Annual Meeting of the Association ofr Computational Linguistics,2019.
[29] Zhao, W. X., dkk.,A survey of large language models,2023. [Online]. Available: arXiv:2303.18223.
[30] Lewis, P., dkk.,Retrieval-augmented generation for knowledge-intensive nlp tasks.Adv. Neural Inf. Process. Syst.,2020,pp.9459-9474.
[31] Maryamah, M., dkk.,*Chatbot*s in academia: A retrieval-augmented generation approach for improved efficient information access.International Conference on Knowledge and Smart Technology.KST,2024,pp.259-264.
[32] LangChainAI,LangChain.[Online].Available: https://github.com/langchainai/langchain
[33] "Analisis hasil survei kebutuhan data 2023,"Badan Pusat Statistik,12 2023[Online]. Available: https://www.bps.go.id/id/publication/2023/12/08/618fb136f757542a3e2f5121/analisis-hasil-surveikebutuhan-data-2023.html
[34] A. Maharana, D.-H. Lee, S. Tulyakov, M. Bansal, F. Barbieri, andY. Fang, "Evaluating very long-term conversational memory of llm agents," arXivpreprint arXiv:2402.17753, 2024.
[35] A. Hodrien dan T. Fernando, "A Review of Post-Study and Post-Task Subjective Questionnaires to Guide Assessment of System Usability," 2021.
[36] J. R. Lewis, "Measuring Perceived Usability: The CSUQ, SUS, and UMUX," Int J Hum Comput Interact, vol. 34, no. 12, hlm. 1148– 1156, Des 2018, doi: 10.1080/10447318.2017.1418805.
[37] J. Sauro dan J. R. Lewis, Quantifying the user experience : practical statistics for user research, 2 ed. Morgan Kaufmann, 2016.
[38] P. B. Brandtzaeg and A. Følstad, "Why people use *Chatbot*s," vol. 10673 LNCS. Springer Verlag, 2017, pp. 377–392.
[39] S. McConnell, Rapid Development: Taming Wild Software Schedules, 1 ed. Microsoft Press, 1996.
[40] W. W. Royce, "Managing The Development of Large Software Systems," 1970. M. E. Khan, "Different approaches to white box testing technique for finding errors," International Journal of Software Engineering and its Applications, vol. 5, no. 3, hlm. 1–14, 2011, doi: 10.5121/ijsea.2011.2404.
[41] S. Nidhra, "Black Box and White Box Testing Techniques – A Literature Review," International Journal of Embedded Systems and Applications, vol. 2, no. 2, hlm. 29–50, Jun 2012, doi: 10.5121/ijesa.2012.2204.
[42] A. Hodrien dan T. Fernando, "A Review of Post-Study and Post-Task Subjective Questionnaires to Guide Assessment of System Usability," 2021. [22] J. R. Lewis, "Measuring Perceived Usability: The CSUQ, SUS, and UMUX," Int J Hum Comput Interact, vol. 34, no. 12, hlm. 1148–1156, Des 2018, doi: 10.1080/10447318.2017.1418805.
[43] J. Sauro dan J. R. Lewis, Quantifying the user experience : practical statistics for user research, 2 ed. Morgan Kaufmann, 2016.