IMPLEMENTASI SISTEM *QUESTION ANSWERING* MENGGUNAKAN METODE *LONG SHORT TERM MEMORY* (LSTM) PADA STUDI KASUS BAHASA SUNDA

Rifal Kurniawan, Teguh Ikhlas Ramadhan, Rudi Hartono
Teknik Informatika, Universitas Perjuangan Tasikmalaya
Jl. Peta No.177, Kahuripan, Kec. Tawang, Tasikmalaya, Jawa Barat 46115
2003010025@unper.ac.id

ABSTRAK

Bahasa Sunda merupakan bagian penting dari warisan budaya yang saat ini menghadapi ancaman kepunahan akibat kontaminasi budaya luar dan penurunan penggunaan di kalangan generasi muda. Latar belakang permasalahan ini menyoroti perlunya upaya serius untuk melestarikan dan memperbarui penggunaan bahasa Sunda agar tetap relevan dan hidup dalam masyarakat modern. Solusi yang ditawarkan dalam penelitian ini adalah implementasi Sistem Question Answering dalam bahasa Sunda, yang dirancang untuk menggali dan melestarikan informasi mendalam tentang budaya Sunda. Penelitian ini fokus pada lima kategori utama: pakakas (peralatan), adat/tradisi, perabot rumah tangga, alat musik/kesenian, dan kadaharan (makanan). Hasil dari penelitian ini diharapkan mampu menyediakan akses mudah dan mendalam terhadap pengetahuan budaya Sunda, serupa dengan yang ditawarkan oleh ChatGPT dalam bahasa Indonesia, tetapi dengan menggunakan bahasa Sunda. Dengan demikian, implementasi ini diharapkan dapat menjadi salah satu langkah konkret dalam menjaga eksistensi dan memperkaya kembali penggunaan bahasa Sunda di tengah arus modernisasi.

Kata kunci : *Bahasa Sunda, Sistem Question Answering , ChatGPT*

1. PENDAHULUAN

GPT (*Generative Pre-trained Transformer*) adalah model bahasa yang dibuat oleh Open AI. Model ini dimaksudkan untuk memahami dan membuat teks bahasa manusia. Untuk memproses dan menghasilkan teks, GPT menggunakan arsitektur transformer yang kuat. ChatGPT adalah aplikasi dari model GPT yang dimaksudkan untuk berkomunikasi dengan manusia melalui teks.[1] Ini dapat digunakan untuk berbagai tujuan, seperti menjawab pertanyaan, memberikan rekomendasi, menghasilkan konten teks, dan melakukan berbagai tugas lain yang terkait dengan komunikasi dengan teks. Ini adalah salah satu contoh penggunaan model GPT dalam aplikasi praktis yang melibatkan interaksi manusia-mesin. ChatGPT merupakan salah satu produk Question Answering yang merupakan teknologi dari kecerdasan buatan yang menggunakan metode LSTM sebagai pengolahan data untuk memberikan respon seperti manusia dalam pemrosesan bahasa alami.[2]

Sistem *Question Answering* merupakan suatu sistem yang didalamnya menggunakan *Natural Language Processing*(NLP) sebagai pemrosesan bahasa alami manusia untuk menggali informasi kemudian secara otomatis menjawab pertanyaan yang telah diajukan.[3] Sistem ini kerap kali dibutuhkan masyarakat seluruh dunia untuk menggali berbagai informasi secara mendalam.

Bahasa sunda termasuk salah satu warisan budaya yang keberadaannya perlu dilestarikan dan dipertahankan karena bahasa sunda kerap digunakan sebagai "Bahasa Indung" oleh kebanyakan masyarakat di daerah Jawa Barat.[4] Data BPS (Badan Pusat Statistik) menunjukkan presentase penutur bahasa daerah semakin menurun pada generasi lebih muda. faktor keluarga, status sosial, tempat tinggal, migrasi, ekonomi, dan multibahasa dalam keluarga adalah penyebab pergeseran bahasa.Hasil penelitian yang dilakukan oleh Balai Bahasa Provinsi Jawa Barat (BBPJB) Kementrian dan Kebudayaan menunjukan bahwa bahasa Sunda terancam punah. Hanya sekitar 40% anak-anak di Jawa Barat yang memahami dan dapat berbahasa Sunda.[5] Maka dari itu, upaya pelestarian dan pembaharuan mengenai bahasa sunda haruslah terus ditingkatkan supaya tidak mengalami kepunahan dan eksistensi bahasa sunda tetap terjaga di kancah nasional terkhusus di daerah Jawa Barat. Implementasi Sistem *Question Answering* menggunakan bahasa sunda merupakan salah satu upaya pelestarian dan penggalian informasi secara mendalam mengenai kekayaan budaya bahasa sunda. Melihat dari fenomena di sekitar kita , orang sunda zaman sekarang yang sudah banyak terkontaminasi budaya-budaya luar sehingga berkurangnya pengetahuan terhadap kekayaan budaya bahasa sunda itu sendiri, yang pada kali ini penulis akan mengimplementasikan sistem *question answering* menggunakan bahasa sunda yang dimana bisa memberikan informasi secara mendalam pada konteks penelitian yang dibagi menjadi 5 kategori yaitu pakakas, adat/tradisi, parabot rumah tangga , alat musik / kesenian , dan kadaharan. Dengan penelitian ini diharapkan informasi dari implementasi yang dilakukan dapat menjadi pengetahuan secara mendalam yang didapatkan dengan mudah seperti halnya ChatGPT namun menggunakan bahasa sunda.

2. TINJAUAN PUSTAKA

2.1. *Natural Language Processing* (NLP)

Pemrosesan bahasa alami (NLP) adalah istilah yang mengacu pada kumpulan teknik komputasi yang dirancang untuk melakukan analisis dan representasi

otomatis bahasa manusia.[6] Namun, untuk mencapai hasil yang sebanding dengan manusia, analisis teks otomatis membutuhkan pemahaman mesin yang jauh lebih dalam tentang bahasa alami. Pengambilan informasi dari internet, agregasi, dan jawaban atas pertanyaan atau sistem question answering adalah beberapa contoh proses pemrosesan bahasa natural (NLP).[7] Proses ini terutama didasarkan pada algoritma yang bergantung pada representasi tekstual halaman web, meskipun NLP juga dapat digunakan sampai batas tertentu. Algoritma ini sangat baik untuk mengambil teks, membaginya menjadi beberapa bagian, memeriksa ejaan dan menganalisis tingkat kata, namun, ia tidak dapat menganalisis kalimat dan paragraf pada tingkat tingkat kata. Oleh karena itu, kemampuan algoritma ini masih sangat terbatas dalam hal menafsirkan kalimat dan mengekstraksi informasi penting dari pertanyaan.

Secara umum, pemrosesan bahasa natural membutuhkan kemampuan simbolis yang tinggi. Kemampuan ini mencakup hal-hal seperti akses dan perolehan karakteristik leksikal, semantik, dan episodik; pembuatan dan penyebaran ikatan dinamis; manipulasi struktur rekursif yang membentuk ikatan; koordinasi berbagai modul pemrosesan dan pembelajaran; dan identifikasi konstruksi bahasa dasar (seperti objek dan tindakan).Untuk beralih dari proses pengolahan bahasa alami (NLP) ke proses yang dikenal sebagai pemahaman bahasa alami, semua keterampilan di atas diperlukan. Saat ini, pendekatan pengolahan bahasa natural bergantung pada representasi sintaksis teks yang juga dikenal sebagai struktur sintaksis dengan menggunakan kata frekuensi kejadian bersama. Batasan algoritma ini adalah mereka hanya dapat memproses informasi dalam teks yang sedang diproses; namun, mereka tidak dapat mempertimbangkan informasi latar belakang tindakan manusia. Kita, misalnya, memahami ketika kita mengatakan bahwa "Sachin Tendulkar adalah batsman yang baik" karena pemahaman kita tentang permainan kriket dan bagaimana "Sachin Tendulkar" berhasil dalam banyak pertandingan kriket. Namun demikian, karena Algoritma saat ini tidak memiliki semua latar belakang ini bersamanya, sehingga pemahamannya terbatas.

Sebagai pengolah teks manusia, kita tidak memiliki batasan seperti itu karena setiap kata yang muncul dalam teks mengaktifkan rangkaian konsep yang terkait secara semantik, episode yang relevan, dan pengalaman sensorik, yang memungkinkan kita untuk menyelesaikan tugas proses pemrosesan bahasa yang rumit. Misalnya, tugas-tugas ini mencakup pelabelan peran semantik dengan cepat dan mudah, serta disambiguasi pengertian kata. Banyak model komputasi baru berusaha mengatasi perbedaan kognitif dengan meniru proses yang dianggap sebagai bagian dari otak manusia dan digunakan untuk memproses bahasa. Metode ini mengandalkan elemen semantik yang tidak dapat diungkapkan secara eksplisit dalam teks. Untuk tujuan teoritis, misalnya untuk studi ilmiah, seperti melihat bagaimana komunikasi linguistik dan karakteristiknya, dan untuk tujuan praktis, misalnya untuk memungkinkan komunikasi manusia-mesin yang efektif.

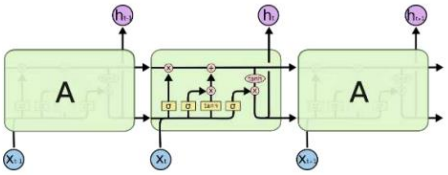
2.2. *Question Answering*

Sistem question answering adalah pencapaian kecerdasan sintetis yang bertahan lama dan kerumitan proses pengolahan bahasa natural.[8] Struktur question answering memungkinkan seseorang mengajukan pertanyaan khusus dalam bahasa alami dan mendapatkan jawaban langsung dan tanggapan singkat. Sekarang, sistem QA ditemukan di mesin pencari seperti Google dan antarmuka percakapan telepon, yang menunjukkan bahwa mereka cukup mahir dalam menanggapi potongan statistik sederhana.Namun, pertanyaan yang lebih menantang biasanya paling mudah untuk dilewati karena pelanggan harus mengembalikan daftar cuplikan yang mereka berikan dan kemudian menelusuri untuk menemukan jawaban pertanyaan mereka.[8]

2.3. *Long Short Term Memory*(LSTM)

Salah satu arsitektur Recurrent Neural Network (RNN) yang berguna untuk menghindari masalah dependensi jangka panjang adalah Long Short-Term Memory (LSTM). Jaringan LSTM terdiri dari empat lapisan yang saling berinteraksi dalam satu neuron: satu lapisan tanh dan tiga lapisan sigmoid. Fungsi aktivasi tanh mengembalikan nilai dalam rentang negatif satu hingga satu, sedangkan fungsi aktivasi sigmoid mengembalikan nilai dalam rentang nol dan satu .

Input gate, forget gate, cell state, dan output gate adalah komponen utama struktur LSTM. Misalkan sequence input x dan hidden state h yang dihasilkan dari lapisan LSTM, maka hubungan antara x dan h yaitu berdasarkan persamaan berikut seperti pada gambar dibawah.



Gambar 1 *Long Short Term Memory* (LSTM)

Jaringan saraf yang dipelajari, atau lapisan jaringan saraf, diwakili oleh kotak berwarna kuning, seperti yang ditunjukkan pada gambar 2.4 1. Vektor transfer, yaitu garis yang membawa seluruh vektor dari keluaran neuron satu ke input neuron lainnya, disebut anak panah tanpa cabang atau garis penggabungan. Lingkaran merah muda adalah operasi pointwise, yaitu operasi yang dilakukan seperti menambah vektor. Proses penggabungan yang menunjukkan rangkuman disebut concatenate. Kopi adalah proses menyalin nilai yang kemudian ditransfer ke tempat lain. Pada langkah waktu t, notasi x_t menunjukkan masukan untuk setiap unit dan nilainya adalah vektor. Notasi h_t menunjukkan state tersembunyi yang dihasilkan oleh

setiap unit pada langkah waktu t dan nilainya adalah vektor. Notasi C_t menunjukkan state sel yang dihasilkan oleh setiap unit pada langkah waktu t dan nilainya berupa vektor.

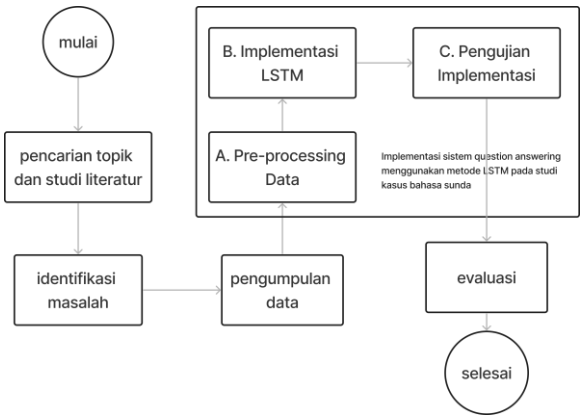
2.4. Bahasa Sunda

Bahasa Sunda adalah cabang Melayu-Polinesia dalam rumpun bahasa Austronesia dengan setidaknya 38 juta penutur, dan merupakan bahasa ibu dengan penutur terbanyak kedua di Indonesia setelah bahasa Jawa. Bahasa ini digunakan di hampir seluruh provinsi Jawa Barat dan Banten, serta di sebagian besar wilayah barat Jawa Tengah, mulai dari Kali Brebes (Sungai Cipamali) di Kabupaten Brebes dan Kali Serayu (Sungai Ciserayu) di Kabupaten Cilacap.[9]

Dengan istilah "undak usuk", bahasa Sunda memiliki tingkatan berbahasa yang hampir tidak ada di bahasa lain. Undak usuk basa dapat dibagi menjadi tiga bagian, yaitu bahasa halus, bahasa sedang, dan bahasa kasar. Undak usuk basa dianggap sebagai tingkatan bahasa, tingkatan, atau tata krama. Selain itu juga, penyebutan dari nama-nama pakakas, parabot, kesenian dan alat musik, kadaharan dan kaulinan memiliki nama dalam bahasa sunda yang unik dan perlu untuk digali informasinya.

a. Pakakas
   Kebanyakan alat yang digunakan orang Sunda untuk bertani. Karena pada masa lalu, pekerjaan utama orang Sunda adalah bercocok tanam dan berternak. Sejak jaman dulu, para karuhun telah diajarkan untuk mengolah sawah, menanam palawija, berkebun, dan berbagai kegiatan lainnya yang sering dilakukan di hutan, seperti berburu.Orang Sunda juga sering dipaksa untuk berperang atau bertempur, sehingga mereka membuat perabot untuk menyerang dan mempertahankan diri, bahkan membunuh lawan. Perkakas berburu biasanya juga digunakan secara tidak langsung dalam berperang. Bahkan menjadi salah satu simbol Sunda, kujang adalah yang paling populer.
b. Parabot
   Parabot rumah tangga sebagai barang-barang rumah tangga tradisional Sunda yang dibuat dengan berbagai bahan seperti bambu, batu, besi, kayu, plastik, tanah liat, aluminium, batok kelapa, dan tembaga. Peralatan ini digunakan untuk berbagai tujuan di rumah. Misalnya, ada dua puluh dua jenis peralatan rumah tangga tradisional yang terbuat dari bambu, yang banyak digunakan karena mudah didapat, mudah dibuat, dan tahan lama. Di sisi lain, hanya satu jenis peralatan rumah tangga yang terbuat dari tembaga, yang paling jarang digunakan dalam peralatan rumah tangga tradisional Sunda.[10]
c. Kesenian dan Alat Musik
   Kesenian termasuk dalam kebudayaan manusia secara keseluruhan karena merupakan representasi dari suatu peradaban yang berkembang sesuai dengan keinginan dan cita-cita yang berpedoman pada nilai-nilai yang berlaku, dan dilakukan melalui aktifitas berkesenian, memberi masyarakat pemahaman tentang jenis keseniannya. Kesenian rakyat adalah salah satu bentuk kreatifitas. Kerajinan orang Sunda biasanya kaya akan nilai, makna, dan kearifan lokal yang terkait dengan rasa syukur terhadap Tuhan, menggambarkan keindahan alam, keramahan masyarakat, interaksi, dan lain-lain. Selain itu, alat musik tradisional sunda yang disebut waditra, juga disebut alat tatabeuhan (tetabuhan) atau instrumen, termasuk bonang, saron, kendang, jenglong, dan gong.[11]
d. Kadaharan
   Pada setiap daerahnya di Indonesia, ada makanan olahan lokal yang resepnya biasanya turun temurun dari generasi ke generasi. Makanan tradisional suku sunda sangat beragam.[12] Masyarakat sunda dapat mengolah satu bahan menjadi berbagai model makanan yang dapat diterima dan disukai oleh semua orang. Seperti halnya bahan dasar sampe, ada banyak jenis makanan tradisional yang menggunakan bahan dasar singkong, seperti combro, misro, oyek, getuk, peyeum, goreng sampeu, dan lain-lain.
e. Tradisi / Adat
   Tradisi dianggap sebagai proses yang berkembang seiring waktu dan tidak statis atau melingkar. Tradisi dapat didefinisikan sebagai warisan yang benar atau warisan masa lalu. Namun, tradisi yang berulang-ulang tidak terjadi secara kebetulan atau disengaja.

3. METODE PENELITIAN



Gambar 2 Metode penelitian

Gambar diatas merupakan metode penelitian yang dilakukan untuk implementasi metode *Long Short Term Memory*(LSTM) pada sistem *question answering* pada studi kasus bahasa sunda. Dimana tahapan-tahapan diatas dilakukan secara berurutan agar dapat tercipta sistem *question answering*. Tahapan tersebut terdiri dari pencarian topik dan studi literatur, identifikasi masalah, pengumpulan data, *pre-processing* data, implementasi LSTM yang meliputi pemilihan model LSTM untuk Memilih arsitektur LSTM yang sesuai dengan kebutuhan sistem question answering, pelatihan model untuk melatih model LSTM dengan dataset yang telah diproses untuk

mengajarkan model bagaimana menjawab pertanyaan dalam bahasa Sunda, optimasi model untuk menyesuaikan hiperparameter dan menggunakan teknik optimasi untuk meningkatkan kinerja model, pengujian implementasi , dan evaluasi.

4. HASIL DAN PEMBAHASAN

4.1. Pengumpulan Data

Proses pengumpulan data dilakukan dengan mengumpulkan kosakata bahasa sunda mengenai 5 kategori penelitian yaitu pakakas, parabot, alat musik/ kasenian, kadaharan dan tradisi/adat. Kosakata tersebut dikembangkan menjadi 12.853 data yang terdiri dari pertanyaan, jawaban serta konteks yang akan menjadi bahan proses penelitian seperti Tabel 1.

Tabel 1 Pengumpulan data

No	Pertanyaan	Jawaban	Konteks
1	leres atanapi henteu alat musik nu utama jang ngiring tari jaipongan namina kecapi	leres	alat musik utama nu ngiring tari jaipong nyaeta kecapi
2	leres atanapi henteu tari anu ngagambarkeun kaendahan bulan purnama namina panon hideung	leres	tari panon hideung nyaeta kasenian tari tradisional sunda anu ngagambarkeun kaendahan bulan purnama
3	leres atanapi henteu tari anu ngagambarkeun gerakan endah nu terinspirasi tina merak namina tari merak	leres	tari merak nyaeta kasenian tari nu ngagambarkeun gerakan endah terinspirasi tina manuk merak
...	….	….	.....
12852	leres atanapi henteu pesta rahayat minangka wujud sukur jeung silih bagi namina pesta laut	henteu	pesta laut nyaeta tradisi nu dilaksanakeun sataun sakali kanggo ngungkapkeun rasa sukur sareng ngadoa kasalametan wanci ek ngalaut
12853	leres atanapi henteu tradisi mandi suci babarengan miceun tingkah polah kurang hade namina pesta laut	henteu	pesta laut nyaeta tradisi nu dilaksanakeun sataun sakali kanggo ngungkapkeun rasa sukur sareng ngadoa kasalametan wanci ek ngalaut

4.2. Preprocessing Data

4.2.1. Inisialisasi tokenizer

Inisialisasi *tokenizer* merupakan tahapan awal untuk melakukan tokenisasi pada dataset dengan memanggil fungsi tokenizer yang dimiliki *tensorflow.* Dengan merubah setiap tulisan menjadi angka atau vektor, di mana angka tersebut adalah nomor khusus untuk setiap kata yang berupa biner (0 dan 1), atau bisa juga berupa vektor dengan nilai yang menunjukkan seberapa sering kata tersebut muncul.[13]

4.2.2. Fit tokenizer pada pertanyaan

Sebelum pertanyaan dilakukan tokenisasi, data hanya berbentuk kolom teks biasa yang didapatkan dari dataset yang tersedia. Proses tqdm menggunakan library pandas dengan deskripsi "*Fitting Tokenizer*" digunakan untuk memantau kemajuan dalam proses pembelajaran komputer, di mana tokenizer diterapkan agar komputer belajar mengenai kata-kata di kolom pertanyaan dalam dataset. Langkah selanjutnya, *tokenizer fit on texts* kedalam data pertanyaan yang merupakan tahapan aktual di mana komputer membaca pertanyaan-pertanyaan tersebut, mengidentifikasi kata-kata yang digunakan, dan memberikan nomor unik (indeks) untuk setiap kata. Setelah tahap tokenisasi, pertanyaan yang tadinya hanya kolom teks biasa sekarang berubah menjadi banyak *array* dari data pertanyaan seperti contoh pada Tabel 2.

Tabel 2 Pertanyaan sebelum dan sesudah tokenisasi

No	Pertanyaan sebelum tokenisasi	Pertanyaan setelah tokenisasi
1	leres atanapi henteu pakakas paranti nakolan paku keur kai jeung sajabana disebut palu	['leres atanapi henteu pakakas paranti nakolan paku keur kai jeung sajabana disebut palu']
2	leres atanapi henteu pakakas paranti motongan kai disebutna ragaji	['leres atanapi henteu pakakas paranti motongan kai disebutna ragaji']
3	leres atanapi henteu pakakas paranti muterkeun baud disebutna obeng	['leres atanapi henteu pakakas paranti muterkeun baud disebutna obeng']
4	leres atanapi henteu pakakas paranti muterkeun mur disebutna konci inggris	['leres atanapi henteu pakakas paranti muterkeun mur disebutna konci inggris']

4.2.3. Konversi Teks Menjadi Urutan Token

Setelah data pertanyaan diubah menjadi bentuk array, selanjutnya digunakan fungsi tqdm dan deskripsi "*Converting to sequences*", yang berguna untuk melihat kemajuan saat komputer mengubah setiap pertanyaan dalam dataset menjadi urutan angka. Kemudian variabel X mewakili fungsi untuk konversi teks pertanyaan menjadi representasi numerik. Dengan memberikan nomor indeks untuk setiap kata berdasarkan pembelajaran sebelumnya oleh *tokenizer*, sistem menciptakan urutan angka yang menyajikan

pertanyaan dalam format yang lebih mudah dipahami oleh komputer seperti pada tabel 3 guna mempersiapkan data untuk langkah-langkah analisis dan pemodelan lebih lanjut.

Tabel 3 Konversi pertanyaan menjadi urutan token

No	Pertanyaan sebelum tokenisasi	Pertanyaan setelah tokenisasi
1	leres atanapi henteu pakakas paranti nakolan paku keur kai jeung sajabana disebut palu	['leres atanapi henteu pakakas paranti nakolan paku keur kai jeung sajabana disebut palu']
2	leres atanapi henteu pakakas paranti motongan kai disebutna ragaji	['leres atanapi henteu pakakas paranti motongan kai disebutna ragaji']
3	leres atanapi henteu pakakas paranti muterkeun baud disebutna obeng	['leres atanapi henteu pakakas paranti muterkeun baud disebutna obeng']
4	leres atanapi henteu pakakas paranti muterkeun mur disebutna konci inggris	['leres atanapi henteu pakakas paranti muterkeun mur disebutna konci inggris']

4.2.4. Padding Urutan Token

Setelah data pertanyaan dijadikan array kemudian diubah lagi menjadi urutan angka unik dari setiap kata, kemudian digunakanlah *tqdm* dan deskripsi "*Padding sequences*", dengan proses ini dapat memantau kemajuan saat komputer melakukan langkah berikutnya. Proses selanjutnya melibatkan fungsi *pad_sequences(X)*, di mana urutan angka sebelumnya (angka yang disimpan dalam variabel X) akan diperlengkap dengan nilai nol, memastikan bahwa semua urutan memiliki panjang yang seragam seperti contoh pada Tabel 4 Pelengkapkan urutan ini berguna ketika sistem menyusun data teks menjadi format yang lebih terstruktur dan seragam. Data yang telah dibuat menjadi berbentuk matriks ini akan mendukung proses analisis atau penggunaan dalam model dengan lebih efektif.

Tabel 4 Padding pada urutan token

No	Pertanyaan sebelum di *padding*	Pertanyaan setelah di *padding*
1	[2, 1, 3, 12, 5, 855, 842, 54, 220, 11, 245, 856, 823]	[0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 2 1 3 12 5 855 842 54 220 11 245 856 823]
2	[2, 1, 3, 12, 5, 246, 220, 181, 112]	[0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 2 1 3 12 5 246 220 181 112]
3	[2, 1, 3, 12, 5, 247, 843, 181, 776]	[0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 2 1 3 12 5 247 843 181 776]
4	[2, 1, 3, 12, 5, 247, 305, 181, 387, 388]	[0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 2 1 3 12 5 247 305 181 387 388]

4.2.5. Pemberian Nilai Pada Kolom Jawaban

Selanjutnya menambahkan proses untuk mempersiapkan data untuk tugas klasifikasi biner, dimana sebelumnya pada kolom jawaban belum terdapat nilai 0 atau pun 1 yang nantinya akan digunakan pada saat pemberian jawaban. Variabel context menampung kolom 'konteks' dari dataset, yang berisi teks konteks atau informasi pendukung. Sedangkan variabel y mengandung label klasifikasi biner, di mana nilai 1 diberikan kepada pertanyaan yang memiliki jawaban 'leres', dan nilai 0 untuk pertanyaan dengan jawaban lainnya. Dengan pendekatan ini, sistem mempersiapkan dua komponen penting dalam pemodelan, yaitu teks konteks sebagai fitur input dan label biner sebagai target output, untuk melatih dan menguji model klasifikasi. Tabel 5 menunjukan jawaban yang dikelompokan menjadi nilai 1 atau 0 .

Tabel 5 Pemberian nilai pada kolom jawaban

No	Jawaban sebelum di beri nilai	Jawaban setelah di beri nilai
1	Leres	1
2	Henteu	0
3	Henteu	0
4	Leres	1

4.2.6. Pembagian Data

Proses selanjutnya memiliki tujuan untuk membagi data menjadi dua set, yakni set pelatihan (*train*) dan set pengujian (*test*), dalam rangka mengembangkan dan mengevaluasi model. Menggunakan tqdm dengan deskripsi "*Splitting data*" memberikan tampilan bar kemajuan saat proses berlangsung. Fungsi *train_test_split* dari pustaka scikit-learn digunakan untuk membagi tiga set data utama: variabel X yang berisi urutan angka dari pertanyaan, variabel context yang berisi teks konteks, dan variabel y yang berisi label biner dari jawaban ('leres' atau bukan). Pembagian dilakukan dengan rasio 80:20 untuk set pengujian dan pelatihan, serta menggunakan nilai acak yang ditentukan (42) untuk memastikan reproduktibilitas hasil. Hasilnya, sistem mendapatkan empat set data: *X_train* dan *y_train* untuk pelatihan, serta *X_test* dan *y_test* untuk pengujian model seperti pada Tabel 6 dan Tabel 7.

Tabel 6 Data latih

No	*X Train* (Urutan Token untuk Data Latih)	*Context Train* (Konteks untuk Data Latih)	*y Train* (Label untuk Data Latih)
1	[0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 2 1 3 12 5 855 842 54 220 11 245 856 823]	cimol nyaeta jajanan khas bandung anu dijieun ...	0
2	[0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 2 1 3 12 5 246 220 181 112]	nyawang bulan nyaeta upacara anu diayakeun man...	0
3	[0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 2 1 3 12 5 247 843 181 776]	kujang nyaeta pakakas ciri khas sunda	0
4	[0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 2 1 3 12 5 247 305 181 387 388]	tauco nyaeta salah sahiji kadaharan khas daera...	0
5	[0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 2 1 3 12 5 34 86 29 91 15 4 389]	batagor nyaeta kadaharan gorengan anu bahan ut...	0

Jumlah data uji: 2571

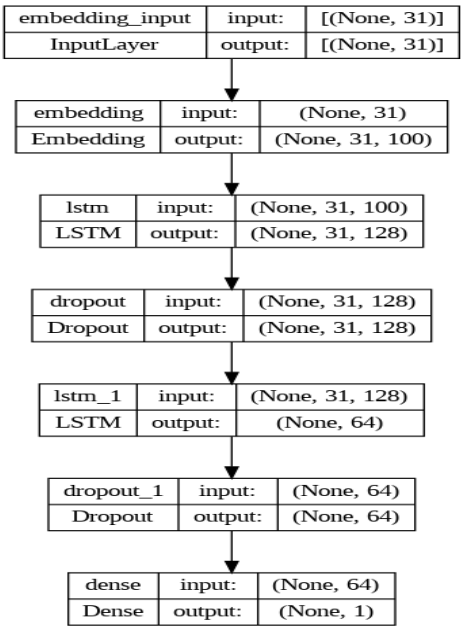
Tabel 7 Data uji

No	*X Test* (Urutan Token untuk Data Uji)	*Context Test* (Konteks untuk Data Uji)	*y Test* (Label untuk Data Uji)
1	[0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 2 1 3 645 91 646 647 181 359 677]	akad nikah nyaeta proses ijab qobul atawa sera...	0
2	[0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 2 1 3 12 820 637 835 15 4 858]	kored nyaeta pakakas paranti miceun jujukutan	1
3	[0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 2 1 3 440 8 5 34 316 158]	abir nyaeta peso badag paranti ngeureutan daging	0
4	[0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 2 1 3 441 12 5 252 72]	damar nyaeta alat paranti nyaangan rohangan ma...	0
5	[0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 2 1 3 132 705 71 30 706 707 4 727]	gusaran nyaeta tradisi ngikir atawa motong wao...	0

Jumlah data latih: 10281

4.3. Implementasi Model Long Short Term Memory

Pada langkah ini, LSTM diimplementasikan pada saat membangun model untuk melakukan pelatihan data yang sebelumnya telah di proses pada tahap pre-processing data. Sebelum membangun model, definisikan terlebih dahulu lokasi yang akan dijadikan penyimpanan model dalam sebuah variabel bernama model_file. Variabel *model_file* ini nantinya akan digunakan untuk menyimpan model jika pelatihan model sudah selesai, dan juga variabel ini digunakan untuk memuat model yang sudah tersimpan ketika pengguna menggunakan kembali sistem ini.



Gambar 3 Implementasi model LSTM

4.3.1. Membuat Sequential Baru

Pada langkah ini, buat model *Sequential* baru dengan fingsi *Sequential*. Model ini adalah salah satu jenis model dalam Keras yang memungkinkan untuk menambahkan lapisan secara berurutan.

4.3.2. Menambahkan Lapisan Embedding

Setelah membuat model, sistem diberikan lapisan embedding yang berfungsi untuk melakukan konversi urutan kata menjadi vektor yang memiliki

representasi numerik yang memiliki nilai setiap kata nya. Parameter-parameter seperti input_dim, output_dim, dan input_length digunakan untuk mengatur ukuran lapisan *embedding*, sedangkan weights diisi dengan matriks *embedding* yang telah dibuat sebelumnya. Dengan *trainable=False* menandakan bahwa bobot lapisan embedding tidak akan diperbarui selama proses pelatihan model.

4.3.3. Menambahkan Lapisan LSTM Pertama

Selanjutnya, sistem ditambahkan lapisan LSTM pertama. Lapisan ini memiliki 128 unit LSTM yang bertujuan untuk memahami hubungan urutan kata dalam pertanyaan dan konteks. Dengan *return_sequences=True* memastikan bahwa lapisan ini memberikan output dari setiap langkah waktu.

4.3.4. Menambahkan Lapisan Dropout Pertama

Setelah lapisan LSTM, sistem ditambahkan lapisan dropout pertama yang berperan dalam mencegah *overfitting* dengan secara acak mengabaikan sebagian *node* di lapisan sebelumnya selama pelatihan. Nilai 0.2 menunjukkan tingkat *dropout* sebesar 20%.

4.3.5. Menambahkan Lapisan LSTM Kedua

Langkah selanjutnya adalah menambahkan lapisan LSTM kedua Lapisan ini memiliki 64 unit LSTM dan bertujuan untuk memproses informasi lebih lanjut dari urutan kata yang telah diolah oleh lapisan sebelumnya.

4.3.6. Menambahkan Lapisan Dropout Kedua

Kemudian, sistem ditambahkan lapisan *dropout* kedua dengan tujuan yang tetap sama, yaitu mencegah *overfitting* dengan mengabaikan sebagian *node* di lapisan sebelumnya secara acak.

4.3.7. Menambahkan Lapisan Dense Terakhir

Langkah terakhir adalah menambahkan lapisan *dense* terakhir .Lapisan ini memiliki satu unit dan menggunakan aktivasi sigmoid, yang menghasilkan output dalam rentang 0 hingga 1 yang dimana output mendekati 1 menunjukkan probabilitas tinggi bahwa pertanyaan sesuai dengan konteksnya. Regularisasi juga diterapkan di sini untuk mencegah *overfitting*.

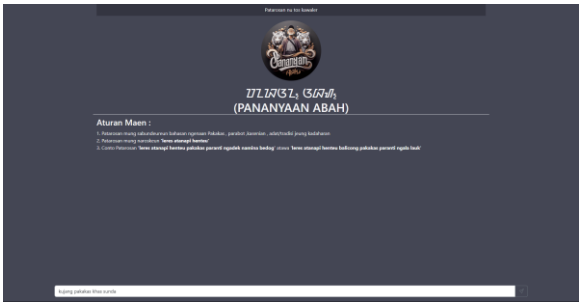
4.3.8. Menambahkan Early Stopping

Early stopping digunakan untuk mencegah overfitting dan menghentikan pelatihan model ketika performa model di atas data validasi tidak lagi meningkat. EarlyStopping adalah fungsi dari Keras yang digunakan untuk memantau performa model selama pelatihan yang memonitoring validation loss. Parameter *validation loss* menentukan metrik yang akan dipantau. Dalam hal ini, sistem akan memantau nilai loss (kerugian) pada data validasi. Ketika nilai loss pada data validasi tidak lagi menurun, pelatihan akan dihentikan. Parameter *patience* adalah jumlah *epochs* (siklus pelatihan) yang akan dipertahankan setelah tidak terjadi peningkatan dalam metrik yang dipantau sebelum pelatihan dihentikan. Pelatihan akan berhenti setelah 5 *epochs* tanpa peningkatan dalam *loss* data validasi.

4.4. Teknis Penggunaan Sistem Question Answering

4.4.1. Input Pertanyaan Ke Sistem *Question Answering*

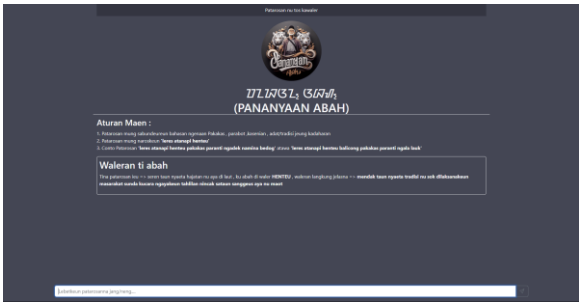
Pertanyaan yang diinputkan oleh pengguna kedalam sistem pada kolom yang tersedia pada tampilan awal sistem dapat berupa kalimat dengan menggunakan bahasa sunda loma atau bahasa sunda sehari-hari dengan konteks bahasan yang sudah ditentukan oleh sistem mengenai 5 kategori yang telah disebutkan diatas seperti gambar 4.



Gambar 4 Input pertanyaan pada sistem

4.4.2. Output Dari Sistem Question Answering

Setelah pengguna mengajukan pertanyaan pada sistem, maka sistem akan memilah dan memilih jawaban yang sesuai dengan konteks yang di pertanyakan oleh pengguna yang nantinya akan memberikan output berupa jawaban antara “leres” atau “henteu” serta pengguna dapat mengetahui konteks yang di pertanyakan dengan jawaban lebih jelas dari sistem seperti pada gambar 5.



Gambar 5 Output dari sistem *question answering*

4.5. Hasil Pengujian Blackbox Interface Sistem

Tabel 8 Hasil pengujian blackbox interface sistem

No	Skenario Pengujian	Uji Kasus	Hasil yang Diharapkan	Hasil Pengujian
1.	Mengosongkan kolom input pertanyaan kemudian klik tombol pesawat ataupun menekan enter.	Input Pertanyaan: (Dikosongkan)	Menampilkan peringatan kolom harus diisi	Sesuai Harapan
2.	Mengisi kolom input dengan pertanyaan	Input pertanyaan: (diisi dengan pertanyaan atau kalimat lainnya)	Sistem membaca pertanyaan dan memberikan output atau jawaban sesuai dengan pertanyaan	Sesuai Harapan
3.	Klik tombol paling atas untuk menampilkan seluruh informasi pertanyaan, jawaban serta konteks yang telah didapat oleh pengguna	-	Menampilkan seluruh informasi dari pertanyaan yang diajukan pengguna serta jawaban dan konteks yang didapat oleh pengguna	Sesuai Harapan
4.	Klik tombol *close* untuk menutup *pop-up* informasi yang tersimpan di sistem	-	Menutup *pop-up* yang ditampilkan	Sesuai Harapan

4.6. Evaluasi

```
# Menghitung akurasi model
accuracy = model.evaluate(X_test, y_test, verbose=0)[1]
print("Akurasi Model:", accuracy)

Akurasi Model: 0.967327892780304
```

Gambar 6 Evaluasi

Tahap evaluasi yaitu untuk mengevaluasi performa model menggunakan data uji *(X_test* dan *y_test*). Dengan menggunakan fungsi *evaluate* akan memberikan hasil evaluasi berupa beberapa metrik performa model, termasuk akurasi. Pada sistem ini menggunakan indeks 1 untuk mengambil nilai akurasi dari hasil evaluasi yang merujuk pada akurasi model. Kemudian nilai akurasi yang diperoleh dari hasil evaluasi disimpan dalam variabel *accuracy*. Sistem *question answering* dengan menggunakan metode LSTM dari 12853 data mendapatkan nilai akurasi model sebesar 0.967 atau 97%.

5. KESIMPULAN DAN SARAN

Implementasi metode *Long Short Term Memory* (LSTM) untuk *Sistem Question Answering* pada studi kasus bahasa sunda ini mendapatkan hasil yang begitu memuaskan pada nilai akurasi model dari hasil pemrosesan 12.853 data yaitu sebesar 97%, namun keakuratan dalam menjawab pertanyaan yang diajukan oleh pengguna masih belum seakurat akurasi model jika pertanyaan yang dimasukan tidak sesuai dengan yang ada pada dataset. Pada sistem question answering kali ini masih terbatas hanya pada jawaban ”leres” atau ”henteu”, sehingga pengguna kurang mengetahui lebih dalam tentang makna dari objek yang diteliti. Penulis menyarankan agar peneliti selanjutnya lebih mengoptimalkan kode untuk pertanyaan serta perbanyak dataset yang akan dijadikan penelitian. 2. Saran selanjutnya dari penulis, peneliti selanjutnya diharapkan bisa memberikan jawaban yang lebih mendetail pada sistem question answering ini tidak hanya terpaku pada jawaban ”leres” atau ”henteu”.

DAFTAR PUSTAKA

[1] W. Suharmawan, “Pemanfaatan Chat GPT Dalam Dunia Pendidikan,” *Education Journal : Journal Educational Research and Development*, vol. 7, no. 2, hlm. 158–166, Agu 2023, doi: 10.31537/ej.v7i2.1248.

[2] Y. B. Kumboro *dkk.*, “PEMANFAATAN CHATGPT SEBAGAI BAHAN REFERENSI KERJA,” 2023. [Daring]. Tersedia pada: https://chat.openai.com.

[3] Y. W. Chandra dan S. Suyanto, “Indonesian chatbot of university admission using a question answering system based on sequence-to-sequence model,” dalam *Procedia Computer Science*, Elsevier B.V., 2019, hlm. 367–374. doi: 10.1016/j.procs.2019.08.179.

[4] H. Nuraeni, “BAHASA TUTUR DALAM KAULINAN URANG LEMBUR DALAM MEMBENTUK KARAKTER ANAK,” 2019.

[5] N. Aini, L. Asri, R. I. Adam, dan B. A. Dermawan, “SPEECH RECOGNITION UNTUK KLASIFIKASI PENGUCAPAN NAMA HEWAN DALAM BAHASA SUNDA MENGGUNAKAN METODE LONG-SHORT TERM MEMORY,” 2023.

[6] M. Zhou, N. Duan, S. Liu, dan H. Y. Shum, “Progress in Neural NLP: Modeling, Learning, and Reasoning,” *Engineering*, vol. 6, no. 3. Elsevier Ltd, hlm. 275–290, 1 Maret 2020. doi: 10.1016/j.eng.2019.12.014.

[7] K. R. Chowdhary, *Fundamentals of artificial intelligence*. Springer India, 2020. doi: 10.1007/978-81-322-3972-7.

[8] A. Agrawal, akhil atri, ayanesh chowdhury, rajeev koneru, kedareswara batchu, dan sai mallavaram, “Question Answering System Using Natural Language Processing,” *International Journal of Research in Engineering, Science and Management*, vol. 4, no. 12, 2021.

[9] R. G. Guntara, A. Nuryadin, dan B. Hartanto, “Pemanfaatan Google Speech to Text Untuk Aplikasi Pembelajaran Kamus Bahasa Sunda Pada Platform Mobile Android,” vol. 4, no. 1,

hlm. 10–19, 2021, doi: 10.31764/justek.vXiY.ZZZ.

[10] Irwansyah dan S. Machdalena, "PENAMAAN PERALATAN RUMAH TANGGA TRADISIONAL PURWA-KARTA JAWA BARAT: KAJIAN ETNOLINGUISTIK," 2022.

[11] H. Maulid dan A. Hasanudin Fauzi, "SEMEN (Sundanese Instrument) : Aplikasi Pengenalan Alat Musik Tradisional Sunda berbasis Augmented Reality," 2018. [Daring]. Tersedia pada: https://www.researchgate.net/publication/344154957

[12] V. Muna Munipati Sukma dan I. Ristri Alkhila, "PENINGKATAN EKSISTENSI MAKANAN TRADISIONAL SUNDA MELALUI OPERASI PASAR BUHUN DI DESA SELAWANGI," *SIWAYANG Journal: Publikasi Ilmiah Bidang Pariwisata, Kebudayaan, dan Antropologi*, vol. 1, no. 2, hlm. 77–84, Jun 2022, doi: 10.54443/siwayang.v1i2.159.

[13] M. Yunus, F. Sains, dan D. Teknologi, "ANALISIS SENTIMEN MENGGUNAKAN METODE DEEP NEURAL NETWORK UNTUK MENGETAHUI RESPON MASYARAKAT INDONESIA TERHADAP DUNIA TRADING Skripsi Oleh : PROGRAM STUDI TEKNIK INFORMATIKA," 2023