

ABSTRAK

Bahasa Madura adalah bahasa ibu bagi penduduk pulau Madura, tetapi saat ini kurang diminati oleh remaja lokal yang lebih sering menggunakan bahasa gaul. Untuk melestarikan bahasa Madura, penelitian ini mengembangkan mesin penerjemah bahasa Madura-Indonesia menggunakan model *Long Short Term Memory* (LSTM) yang bergantung pada arsitektur *Encoder-Decoder*. Model ini melalui beberapa tahap pemrosesan data seperti *tokenizing* dan normalisasi, dengan data dipisahkan menjadi 80% untuk pelatihan dan 20% untuk pengujian. Sistem penerjemah ini diimplementasikan menggunakan *Flask* sebagai server untuk menghubungkan aplikasi Android dengan model terjemahan. Berdasarkan hasil *BLEU-score*, parameter seperti *Batch Size* dan *Epochs* memiliki dampak signifikan terhadap kinerja model. Percobaan pertama dengan *Batch Size* 126 dan *Epochs* 36 menunjukkan *BLEU-score* 0.3095 untuk 1-grams, namun menurun untuk *n-grams* yang lebih tinggi. Percobaan kedua dengan *Epochs* 50 menunjukkan penurunan *BLEU-score* menjadi 0.2534 untuk 1-grams. Percobaan dengan *Batch Size* 256 menghasilkan *BLEU-score* yang sangat rendah, sedangkan *Batch Size* 64 menunjukkan hasil terbaik dengan *BLEU-score* 0.3139 pada *Epochs* 36 dan tetap stabil pada *Epochs* 50. Percobaan-percobaan ini menunjukkan bahwa model mampu memberikan terjemahan akurat dan efisien dengan antarmuka yang responsif, meskipun semua model menunjukkan kecenderungan *overfitting* pada berbagai skenario uji coba.

Kata kunci : bahasa, mesin penerjemah, LSTM, aplikasi android

ABSTRACT

The Madurese language is the native language of the inhabitants of Madura Island. However, it is currently less favored by local teenagers, who prefer to use colloquial language. To preserve the Madurese language, this study developed a Madurese-Indonesian translation machine using a Long Short Term Memory (LSTM) model, which relies on the Encoder-Decoder architecture. The model underwent several data processing stages, such as tokenizing and embedding, with the data divided into 80% for training and 20% for testing. The translation system was implemented using Flask as the server to connect an Android application with the translation model hosted on GitHub. According to the BLEU-score results, parameters such as Batch Size and Epochs significantly impact the model's performance. The first experiment, with Batch Size 126 and 36 Epochs, showed a BLEU-score of 0.3095 for 1-grams, but the score decreased for higher n-grams. The second experiment, with 50 Epochs, showed a further decrease in BLEU-score to 0.2534 for 1-grams. The experiment with Batch Size 256 yielded very low BLEU-scores, whereas Batch Size 64 produced the best results, with a BLEU-score of 0.3139 at 36 Epochs and remained stable at 50 Epochs. These experiments indicate that the model can provide accurate and efficient translations with a responsive interface, despite all models showing a tendency to overfit in various test scenarios.

Keywords: *Madurese language, machine translation, LSTM, neural network, Flask, BLEU-score, overfitting, Batch Size, Epochs.*

DAFTAR ISI

HALAMAN JUDUL	i
HALAMAN PENGESAHAN	ii
HALAMAN PERNYATAAN ORISINALITAS.....	iii
KATA PENGANTAR.....	iv
ABSTRAK.....	vi
ABSTRACT	vii
DAFTAR ISI	viii
DAFTAR TABEL	xi
DAFTAR GAMBAR.....	xii
BAB I PENDAHULUAN.....	1
1.1 Latar Belakang.....	1
1.2 Rumusan Masalah.....	9
1.3 Pertanyaan Penelitian	9
1.4 Tujuan dan Manfaat Penelitian.....	10
1.5 Batasan Masalah.....	10
1.6 Sistematika Penulisan.....	11
BAB II DASAR TEORI.....	12
2.1 Landasan Teori	12
2.1.1 Bahasa Daerah.....	12
2.1.2 Bahasa Madura	12
2.1.3 <i>Machine Translator</i> (Mesin Penerjemah)	12
2.1.4 Neural Machine Translation (NMT).....	13
2.1.5 NLP (Natural Language Processing).....	14
2.1.6 <i>Deep learning</i>	15
2.1.7 Fungsi Aktivasi	17
2.1.7.1 Fungsi Aktivasi Sigmoid Biner	17
2.1.7.2 Fungsi Aktivasi Sigmoid Bipolar.....	18

2.1.8	RNN (<i>Recurrent Neural Network</i>).....	19
2.1.8	<i>Sequential to Sequential</i>	23
2.1.8.1	Encoder	23
2.1.8.2	Decoder	23
2.1.9	<i>Word2vec</i>	24
2.1.10	LSTM (<i>Long Short Term Memory</i>).....	28
2.1.11	<i>Bleu-score</i>	33
2.2	Penelitian Terdahulu.....	34
BAB III METODE PENELITIAN		42
3.1	Alur Penelitian.....	42
3.2	Uraian Penelitian	43
3.2.1	<i>Dataset</i>	43
3.2.2	<i>Preprocessing Data</i>	44
3.2.2.1	Tokenizing.....	44
3.2.3	<i>Embedding</i>	47
3.2.4	Pemodelan dan Pelatihan	48
3.2.5	Evaluasi.....	51
3.2.6	Pengimplementasian Sistem.....	51
3.3	Skenario Pengujian.....	53
BAB IV HASIL PENELITIAN DAN PEMBAHASAN.....		56
4.1	Lingkungan Uji Coba	56
4.2	Implementasi Kode Program.....	59
4.2.1	<i>Input Dataset</i>	60
4.2.2	Pemrosesan <i>Dataset</i>	63
4.2.3	<i>Tokenizing Dataset</i>	66
4.2.4	<i>Encode-Decode</i>	68

4.2.5 Mempersiapkan Data <i>Source</i> Dan Data Target	71
4.2.6 Menyiapkan Data Latih Dan Data Uji	72
4.2.7 Pemodelan Dan Menyimpan Hasil Model.....	73
4.2.8 Hasil Prediksi Terjemahan.....	76
4.2.9 Hasil Evaluasi BLEU <i>Score</i>	78
4.3 Kesalahan Pada <i>Dataset</i> Yang Digunakan	80
4.4 Statistik <i>Dataset</i> Yang Digunakan	81
4.4.1 Hasil <i>Preprocessing</i> Dan Normalisasi <i>Dataset</i>	86
4.4.1.1 Distribusi Panjang Kalimat	87
4.4.1.2 Frekuensi Kata	93
4.4.1.3 Analisis Perubahan Kalimat Setelah Pemrosesan.....	98
4.4.2 Statistik Hasil Tokenisasi <i>Dataset</i>	99
4.5 <i>Preprocessing</i> Data dengan <i>Encoding</i> dan <i>Decoding</i>	101
4.6 Pembagian Data Training Dan Data Testing.....	104
4.7 Pembentukan Model.....	106
4.8 Hasil Evaluasi Model.....	110
4.9 <i>Prototype</i> Tampilan Hasil Pengimplementasian Aplikasi.....	117
4.9.1 Pengimplementasian <i>Flask</i>	118
4.9.2 Aplikasi <i>Android</i>	119
BAB V PENUTUP	122
5.1 Kesimpulan.....	122
5.2 Saran.....	123
DAFTAR PUSTAKA.....	125
DAFTAR RIWAYAT HIDUP	128

DAFTAR TABEL

Tabel 2.1 Penelitian Terkait Mengenai LSTM.....	36
Tabel 3.1 Contoh <i>Dataset</i> yang digunakan.....	43
Tabel 3.3 Contoh data yang sudah melewati tahapan <i>Tokenizing</i>	46
Tabel 3.4 Data Training	48
Tabel 3.5 Data <i>Testing</i>	49
Tabel 3.6 Model dan <i>Layer</i> LSTM	49
Tabel 3.7 Skenario pengujian	53
Tabel 4.1 Spesifikasi perangkat keras yang digunakan.....	56
Tabel 4.2 Perangkat lunak yang digunakan.....	57
Tabel 4.3 Statistik kalimat sumber dan target	81
Tabel 4.4 10 kata yang paling sering muncul.....	83
Tabel 4.5 Frekuensi Huruf Beaksen	83
Tabel 4.6 Panjang Dokumen.....	84
Tabel 4.7 Hasil Visualisasi Plot Untuk Masing-Masing Bahasa	89
Tabel 4.8 Tabel Hasil Frekuensi Kata Pada Bahasa Madura Dan Bahasa Indonesia.....	94
Tabel 4.9 Contoh Kalimat Sebelum Dan Sesudah Pemrosesan	98
Tabel 4.10 Statistik Hasil Tokenisasi.....	99
Tabel 4.11 Contoh Hasil <i>Encode</i>	103
Tabel 4.12 Struktur Model.....	109
Tabel 4.13 Hasil Akurasi Pada Masing-Masing Skenario Uji Coba.....	112

DAFTAR GAMBAR

Gambar 2.1 Model <i>Encoder-Decoder</i>	13
Gambar 2.2 Ilustrasi <i>Deep learning</i> [12].....	16
Gambar 2.3 Fungsi Aktivasi Sigmoid Biner[13]	17
Gambar 2.4 Fungsi Aktivasi Sigmoid Bipolar[13]	18
Gambar 2.5 Arsitektur RNN dengan Satu <i>Layer</i> dan Banyak <i>Layer</i> [12].....	19
Gambar 2.6 Pemrosesan RNN[12]	20
Gambar 2.7 Struktur RNN[12]	21
Gambar 2.8 Model <i>SeqtoSeq</i> [12].....	23
Gambar 2.9 CBOW Satu Kata Konteks[15]	24
Gambar 2.10 CBOW Multipel Kata Konteks[15]	25
Gambar 2.11 <i>Skip-gram</i> Arsitektur[15].....	25
Gambar 2.12 <i>Long Short Term Memory</i> (LSTM)[12].....	28
Gambar 2.13 Gerbang Forget LSTM[12]	29
Gambar 2.14 Gerbang <i>Input</i> LSTM[12]	30
Gambar 2.15 Gerbang Update LSTM[12]	31
Gambar 2.16 Gerbang <i>Output</i> LSTM[12]	32
Gambar 3.1 Alur Penelitian	42
Gambar 3.2 <i>Flowchart tokenizing</i>	45
Gambar 3.4 Desain Arsitektur	51
Gambar 3.5 <i>Main Page Interface</i> Aplikasi Mesin Penerjemah	52
Gambar 4.1 Hasil Distribusi Panjang Kalimat	88
Gambar 4.2 Hasil Distribusi Panjang Kalimat Bahasa Indonesia.....	88
Gambar 4.3 Hasil Perbandingan Frekuensi Kata Bahasa Madura	93
Gambar 4.4 Hasil Perbandingan Frekuensi Kata Bahasa Madura	94
Gambar 4.5 Hasil Implementasi Bentuk Model	110
Gambar 4. 6 Hasil <i>Bleu-score</i> Untuk Masing-Masing Model	112
Gambar 4.7 Hasil Visualisasi Plot Skenario Uji Coba.....	114
Gambar 4.8 Hasil Tampilan Aplikasi	121