

# **IDENTIFIKASI BAHASA PADA TEKS MENGGUNAKAN METODE *LONG SHORT TERM MEMORY (LSTM)***

Diajukan Sebagai Syarat Untuk Menyelesaikan  
Pendidikan Program Strata-1 Pada  
Jurusan Teknik Informatika



Oleh :

Sheva Satrian

NIM : 09021282025081

**Jurusan Teknik Informatika**

**FAKULTAS ILMU KOMPUTER UNIVERSITAS SRIWIJAYA**

**2023**

## LEMBAR PENGESAHAN SKRIPSI

### IDENTIFIKASI BAHASA PADA TEKS MENGGUNAKAN METODE *LONG SHORT TERM MEMORY (LSTM)*

Oleh :

Sheva Satrian

NIM : 09021282025081

Palembang, 20 Desember 2023

Pembimbing I



Alvi Syahrini Utami, M.Kom.  
NIP. 197812222006042003

Pembimbing II,



Junia Kurniati, M.Kom.  
NIK. 1671046606890018

Mengetahui,

Ketua Jurusan Teknik Informatika



Alvi Syahrini Utami, M.Kom.  
NIP. 197812222006042003

## TANDA LULUS UJIAN KOMPREHENSIF

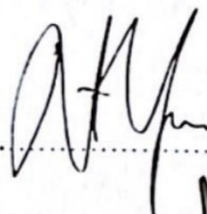
Pada hari Selasa tanggal 19 Desember 2023 telah dilaksanakan ujian Komprehensif skripsi oleh Jurusan Teknik Informatika Fakultas Ilmu Komputer Universitas Sriwijaya.

Nama : Sheva Satrian  
NIM : 09021282025081  
Judul : Identifikasi Bahasa Pada Teks Menggunakan Metode *Long Short Term Memory (LSTM)*

dan dinyatakan **LULUS**.

1. Ketua Penguji

Novi Yusliani, M.T.  
NIP. 198211082012122001

  
.....


2. Penguji

Dr. Abdiansah, S.Kom., M.Cs.  
NIP. 198410012009121005

  
.....

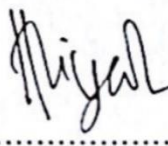
3. Pembimbing 1

Alvi Syahrini Utami, M.Kom.  
NIP. 197812222006042003

  
.....

4. Pembimbing 2

Junia Kurniati, M.Kom.  
NIK. 1671046606890018

  
.....

Mengetahui,  
Ketua Jurusan Teknik Informatika



Alvi Syahrini Utami, M.Kom.  
NIP: 197812222006042003



## HALAMAN PERNYATAAN BEBAS PLAGIAT

Yang bertanda tangan dibawah ini:

Nama : Sheva Satrian  
NIM : 09021282025081  
Program Studi : Teknik Informatika Reguler  
Judul : Identifikasi Bahasa Pada Teks Menggunakan  
Metode *Long Short Term Memory*

### Hasil Pengecekan Software iThenticate/Turnitin: 13%

Menyatakan bahwa laporan skripsi saya merupakan hasil karya sendiri dan bukan hasil penjiplakan/plagiat. Apabila ditemukan unsur penjiplakan atau plagiat dalam laporan tugas akhir ini, maka saya bersedia menerima sanksi akademik dari Universitas Sriwijaya

Demikian pernyataan ini saya buat dengan sebenarnya dan tidak ada paksaan dari siapapun.

Palembang, 7 Desember 2023

Penulis,



Sheva Satrian  
NIM. 09021282025081

## **MOTTO DAN PERSEMBAHAN**

“Jika merasa gagal dalam mencapai mimpi, jangan khawatir mimpi-mimpi lain bisa diciptakan, jangan menyerah tetaplah berjuang bangkit dari keterpurukan karena kita disini petarung untuk kehidupan yang keras ini” – Windah Basudara

Kupersembahkan karya tulis ini kepada:

- Allah SWT
- Orang Tua dan Keluargaku
- Teman-teman penulis
- Universitas Sriwijaya

## ABSTRACT

*Language is the main communication tool used by humans, with the diversity of languages that exist in the world reflecting the cultural diversity and identity of a language. In this context, language identification is important for the development of communication technology and information processing. This research focuses on language identification in text by utilizing Long Short Term Memory method and Word2vec as Word Embedding method to produce effective results from text. The main objective of this research is to develop a system that is able to recognize and classify language in text with high accuracy. The dataset used in this research consists of 10,000 text data, which includes 10 different language label classes with 1000 data each including Arabic, Chinese, Dutch, English, French, Indonesian, Japanese, Korean, Russian, Spanish. The total dataset is divided into 80% training data and 20% test data, to determine the hyperparameters used in the study by searching using the random search method. After the process, the best hyperparameter results were obtained for the LSTM model with a dropout configuration of 0.3, batch size 32, hidden unit 64, recurrent dropout 0.2 and epoch 15. Based on this research, by evaluating using the confusion matrix table, the average value of evaluation metrics such as precision 0.9859, recall 0.9855 and f1-score 0.9856 and getting an accuracy value of 0.9856.*

*Keywords: Language Identification, Confusion Matrix, Long Short Term Memory, Word Embedding, Text, Word2Vec*

## ABSTRAK

Bahasa adalah alat komunikasi utama yang digunakan oleh manusia, dengan keragaman bahasa yang ada di dunia mencerminkan keanekaragaman budaya dan identitas suatu bahasa. Dalam konteks ini, identifikasi bahasa penting untuk pengembangan teknologi komunikasi dan pengolahan informasi. Penelitian ini berfokus pada identifikasi bahasa dalam teks dengan memanfaatkan metode *Long Short Term Memory* dan *Word2vec* sebagai metode *Word Embedding* untuk menghasilkan hasil yang efektif dari teks. Tujuan utama dari penelitian ini adalah untuk mengembangkan sistem yang mampu mengenali dan mengklasifikasikan bahasa pada teks dengan akurasi tinggi. Dataset yang digunakan dalam penelitian ini terdiri dari 10.000 data teks, yang mencakup 10 kelas label bahasa yang berbeda dengan masing-masing 1000 data diantaranya *Arabic, Chinese, Dutch, English, French, Indonesian, Japanese, Korean, Russian, Spanish*. Total dari dataset tersebut dibagi menjadi data latih sebanyak 80% dan data uji sebanyak 20%, untuk menentukan *hyperparameter* yang digunakan pada penelitian yaitu dengan mencari menggunakan metode *random search*. Setelah proses tersebut, diperoleh hasil *hyperparameter* terbaik untuk model LSTM dengan konfigurasi *dropout* 0.3, *batch size* 32, *hidden unit* 64, *recurrent dropout* 0.2 dan *epoch* 15. Berdasarkan penelitian tersebut, dengan evaluasi menggunakan tabel *confusion matrix* didapatkan nilai rata-rata metrik evaluasi seperti *precision* 0.9859, *recall* 0.9855 dan *f1-score* 0.9856 serta mendapat nilai *accuracy* sebesar 0.9855.

Kata Kunci : Identifikasi Bahasa, *Confusion Matrix*, *Long Short Term Memory*, *Word Embedding*, Teks, *Word2Vec*

## KATA PENGANTAR

Puji dan syukur penulis ucapkan kepada Allah SWT atas berkat dan anugerah-Nya sehingga penulis dapat menyelesaikan skripsi ini dengan judul “Identifikasi Bahasa Pada Teks Menggunakan Metode *Long Short Term Memory*”. Skripsi ini dibuat sebagai salah satu syarat untuk menyelesaikan program Sarjana (S1) Jurusan Teknik Informatika Fakultas Ilmu Komputer Universitas Sriwijaya.

Dalam menyelesaikan skripsi ini penulis menerima banyak dukungan dan bantuan dari berbagai pihak, baik yang diberikan secara langsung maupun tidak langsung. Atas hal tersebut, penulis ingin menyampaikan rasa terima kasih yang sebesar-besarnya kepada:

1. Keluarga tercinta yang telah memberikan dukungan dan doa kepada penulis hingga dapat menyelesaikan skripsi ini.
2. Bapak Jaidan Jauhari, S.Pd., M.T. (alm) selaku Dekan Fakultas Ilmu Komputer.
3. Ibu Alvi Syahrini Utami, M.Kom. selaku pembimbing skripsi sekaligus Ketua Jurusan Teknik Informatika.
4. Ibu Junia Kurniati, M.Kom. selaku Dosen Pembimbing yang telah memberikan arahan, masukan, kritik dan saran kepada saya dalam menyelesaikan tugas akhir ini.
5. Bapak Kanda Januar Miraswan, M.T. selaku Dosen Pembimbing Akademik yang telah memberikan bimbingan, arahan, dan motivasi dalam proses perkuliahan.



6. Semua Dosen dan Staf Jurusan Teknik Informatika di Fakultas Ilmu Komputer Universitas Sriwijaya, yang telah memberikan pengetahuan dan dukungan kepada penulis selama perkuliahan.
7. Sahabat seperjuangan dan teman-teman Teknik Informatika Reguler B 2020.
8. Serta semua pihak yang telah mendukung dalam penyelesaian skripsi ini.

Penulis menyadari bahwa dalam penulisan skripsi ini masih ada kekurangan karena batasan pengetahuan dan pengalaman. Oleh karena itu, penulis berharap kritik dan saran yang membangun akan menyempurnakan skripsi ini dan agar dapat bermanfaat bagi semua pihak.

Palembang, 19 Desember 2023

Sheva Satrian

## DAFTAR ISI

	Halaman
LEMBAR PENGESAHAN SKRIPSI .....	ii
TANDA LULUS UJIAN KOMPREHENSIF.....	iii
HALAMAN PERNYATAAN BEBAS PLAGIAT .....	iv
MOTTO DAN PERSEMBAHAN .....	v
ABSTRACT .....	vi
ABSTRAK .....	vii
KATA PENGANTAR .....	viii
DAFTAR ISI.....	x
DAFTAR TABEL.....	xiv
DAFTAR GAMBAR .....	xv
DAFTAR ISTILAH .....	xvii
BAB I PENDAHULUAN .....	I-1
1.1    Pendahuluan .....	I-1
1.2    Latar Belakang .....	I-1
1.3    Rumusan Masalah .....	I-3
1.4    Tujuan Penelitian.....	I-3
1.5    Manfaat Penelitian.....	I-3
1.6    Batasan Masalah .....	I-3
1.7    Sistematika Penulisan.....	I-4
1.8    Kesimpulan.....	I-5
BAB II KAJIAN LITERATUR .....	II-1
2.1    Pendahuluan .....	II-1
2.2    Landasan Teori .....	II-1
2.2.1 <i>Natural Language Processing</i> .....	II-1
2.2.2 <i>Language Identification</i> .....	II-2
2.2.3 <i>Preprocessing Text</i> .....	II-3
2.2.4 <i>Word2vec</i> .....	II-6
2.2.5 <i>Artificial Neural Network</i> .....	II-7
2.2.6 <i>Long Short Term Memory (LSTM)</i> .....	II-8

2.2.7	<i>Confusion Matrix</i> .....	II-12
2.2.8	<i>Rational Unified Process</i> .....	II-14
2.3	Penelitian Lain yang Relevan .....	II-15
2.4	Kesimpulan.....	II-16
BAB III METODOLOGI PENELITIAN.....		III-1
3.1	Pendahuluan .....	III-1
3.2	Pengumpulan Data .....	III-1
3.2.1	Jenis Data .....	III-1
3.2.2	Sumber Data .....	III-1
3.2.3	Metode Pengumpulan Data.....	III-2
3.3	Tahapan Penelitian .....	III-2
3.3.1	Mengumpulkan Data .....	III-3
3.3.2	Menentukan Kerangka Kerja Penelitian .....	III-3
3.3.3	Menentukan Kriteria Pengujian.....	III-10
3.3.4	Menentukan Format Data Pengujian .....	III-10
3.3.5	Menentukan Alat Bantu Penelitian.....	III-12
3.3.6	Melakukan Pengujian Penelitian.....	III-13
3.3.7	Melakukan Analisis dan Menarik Kesimpulan.....	III-13
3.4	Metode Pengembangan Perangkat Lunak .....	III-14
3.4.1	Fase Insepsi .....	III-14
3.4.2	Fase Elaborasi .....	III-14
3.4.3	Fase Konstruksi.....	III-15
3.4.4	Fase Transisi .....	III-15
3.5	Manajemen Proyek Penelitian .....	III-15
3.6	Kesimpulan.....	III-16
BAB IV PENGEMBANGAN PERANGKAT LUNAK .....		IV-1
4.1	Pendahuluan .....	IV-1
4.2	Fase Insepsi .....	IV-1
4.2.1	Pemodelan Bisnis .....	IV-1
4.2.2	Kebutuhan Sistem .....	IV-4
4.2.3	Analisis dan Desain.....	IV-5
4.2.3.1	Analisis Kebutuhan Perangkat Lunak.....	IV-5
4.2.3.2	Analisis Data.....	IV-6
4.2.3.3	Analisis <i>preprocessing</i> .....	IV-6

4.2.3.4	Analisis Proses Klasifikasi .....	IV-27
4.2.4	Implementasi .....	IV-28
4.2.4.1	<i>Use Case</i> .....	IV-28
4.2.4.2	Tabel Definisi Pengguna.....	IV-29
4.2.4.3	Tabel Definisi <i>Use Case</i> .....	IV-30
4.2.4.4	Tabel Skenario <i>Use Case</i> .....	IV-31
4.3	Fase Elaborasi.....	IV-36
4.3.1	Pemodelan Bisnis .....	IV-36
4.3.2	Perancangan Data.....	IV-36
4.3.3	Perancangan Antar Muka.....	IV-37
4.3.4	Kebutuhan Sistem .....	IV-39
4.3.5	<i>Activity Diagram</i> .....	IV-40
4.3.6	<i>Sequence Diagram</i> .....	IV-43
4.4	Fase Konstruksi .....	IV-45
4.4.1.	Kebutuhan Sistem .....	IV-45
4.4.2.	<i>Class Diagram</i> .....	IV-45
4.4.3.	Implementasi .....	IV-46
4.4.3.1	Implementasi Kelas.....	IV-46
4.4.3.2	Implementasi Antarmuka.....	IV-47
4.5	Fase Transisi.....	IV-50
4.5.1	Pemodelan Bisnis .....	IV-51
4.5.2	Rencana Pengujian .....	IV-51
4.5.3	Implementasi .....	IV-53
4.5.3.1	Pengujian <i>Use Case Upload Dataset</i> .....	IV-53
4.5.3.2	Pengujian <i>Use Case Preprocessing Dataset</i> .....	IV-53
4.5.3.3	Pengujian <i>Use Case Melakukan Proses Word2vec</i> .....	IV-54
4.5.3.4	Pengujian <i>Use Case Training Model LSTM</i> .....	IV-55
4.5.3.5	Pengujian <i>Use Case Melakukan Identifikasi Bahasa</i> .....	IV-56
4.6	Kesimpulan.....	IV-56
BAB V HASIL DAN ANALISIS PENELITIAN.....		V-1
5.1	Pendahuluan .....	V-1
5.2	Data Hasil Penelitian .....	V-1
5.2.1	Konfigurasi Percobaan .....	V-1
5.2.2	Data Hasil Konfigurasi.....	V-2

5.3	Analisis Hasil Penelitian .....	V-4
5.4	Kesimpulan.....	V-10
BAB VI KESIMPULAN DAN SARAN .....		VI-1
6.1	Pendahuluan .....	VI-1
6.2	Kesimpulan.....	VI-1
6.3	Saran .....	VI-1
DAFTAR PUSTAKA .....		xix
LAMPIRAN.....		xxiii



## DAFTAR TABEL

	Halaman
Tabel III- 1 Rancangan Tabel Pengujian <i>Confusion Matrix</i> .....	III-10
Tabel III- 2 Tabel Hasil <i>Confusion Matrix</i> .....	III-11
Tabel III- 3 Rancangan Tabel Hasil Analisis Identifikasi .....	III-13
Tabel IV- 1 Kebutuhan Fungsional Perangkat Lunak.....	IV-4
Tabel IV- 2 Kebutuhan Non-Fungsional Perangkat Lunak .....	IV-5
Tabel IV- 3 Contoh Teks .....	IV-6
Tabel IV- 4 Hasil Proses <i>Cleaning</i> data.....	IV-9
Tabel IV- 5 Hasil Proses Tokenisasi data .....	IV-12
Tabel IV- 6 Hasil Proses mengubah teks menjadi Token .....	IV-18
Tabel IV- 7 Hasil Proses <i>Padding</i> .....	IV-22
Tabel IV- 8 Contoh <i>Word Vector</i> “di” dengan Dimensi 400.....	IV-27
Tabel IV- 9 Rentang Nilai <i>Hyperparameter</i> .....	IV-28
Tabel IV- 10 Definisi Pengguna.....	IV-29
Tabel IV- 11 Definisi <i>Use Case</i> .....	IV-30
Tabel IV- 12 Skenario <i>Upload</i> Dataset .....	IV-31
Tabel IV- 13 Skenario <i>Preprocessing</i> Dataset .....	IV-32
Tabel IV- 14 Skenario Melakukan Proses <i>Word2vec</i> .....	IV-33
Tabel IV- 15 Skenario <i>Training</i> model LSTM .....	IV-34
Tabel IV- 16 Skenario Melakukan identifikasi bahasa.....	IV-35
Tabel IV- 17 Implementasi Kelas .....	IV-46
Tabel IV- 18 Rencana Pengujian <i>Upload</i> Dataset .....	IV-51
Tabel IV- 19 Rencana Pengujian <i>Preprocessing text</i> .....	IV-51
Tabel IV- 20 Rencana Pengujian <i>Word2vec</i> .....	IV-52
Tabel IV- 21 Rencana Pengujian <i>Training</i> model LSTM .....	IV-52
Tabel IV- 22 Rencana Pengujian Proses Pengujian Identifikasi Bahasa .....	IV-52
Tabel IV- 23 Hasil Pengujian <i>Use Case Upload</i> Dataset.....	IV-53
Tabel IV- 24 Hasil Pengujian <i>Use Case Preprocessing</i> Dataset .....	IV-54
Tabel IV- 25 Hasil Pengujian <i>Use Case</i> Melakukan Proses <i>Word2vec</i> .....	IV-54
Tabel IV- 26 Hasil Pengujian <i>Use Case Training</i> Model LSTM.....	IV-55
Tabel IV- 27 Hasil Pengujian Melakukan Identifikasi Bahasa.....	IV-56
Tabel V- 1 Konfigurasi <i>Hyperparameter</i> .....	V-1
Tabel V- 2 Tabel <i>Confusion Matrix</i> .....	V-2
Tabel V- 3 Tabel Hasil <i>Confusion Matrix</i> .....	V-3
Tabel V- 4 Hasil Percobaan <i>Training</i> Model LSTM .....	V-5
Tabel V- 5 Hasil Performa <i>Training</i> Model LSTM .....	V-5
Tabel V- 6 Sampel hasil prediksi identifikasi bahasa .....	V-8

## DAFTAR GAMBAR

	Halaman
Gambar II- 1 Contoh Proses <i>Case Folding</i> .....	II- 4
Gambar II- 2 Contoh Proses <i>Trim Text</i> .....	II-4
Gambar II- 3 Contoh Proses <i>Remove Punctuations, Special Characters, and Double Whitespace</i> .....	II-4
Gambar II- 4 Contoh Proses <i>Number Removal</i> .....	II-5
Gambar II- 5 Contoh Proses <i>Tokenize</i> .....	II-5
Gambar II- 6 Contoh Proses <i>Stemming</i> .....	II-6
Gambar II- 7 Arsitektur <i>Artificial Neural Network</i> (Windarto et al., 2018).....	II-8
Gambar II- 8 Arsitektur <i>Long Short Term Memory</i> (Le et al., 2019).....	II-9
Gambar II- 9 Model Confusion Matrix (Grandini et al., 2020).....	II-12
Gambar III- 1 Rincian Langkah-langkah Kegiatan Penelitian.....	III-2
Gambar III- 2 Kerangka Kerja Penelitian.....	III-3
Gambar III- 3 <i>Flowchart</i> Tahapan Text Preprocessing .....	III-6
Gambar III- 4 <i>Flowchart</i> model LSTM.....	III-8
Gambar IV- 1 Diagram <i>Use Case</i> .....	IV-29
Gambar IV- 2 Rancangan Antarmuka.....	IV-37
Gambar IV- 3 Rancangan Antarmuka <i>Upload Dataset</i> .....	IV-37
Gambar IV- 4 Rancangan Antarmuka <i>Preprocessing Dataset</i> .....	IV-38
Gambar IV- 5 Rancangan Antarmuka <i>Word2vec</i> .....	IV-38
Gambar IV- 6 Rancangan Antarmuka <i>Training Model LSTM</i> .....	IV-39
Gambar IV- 7 Rancangan Antarmuka <i>Prediksi Identifikasi Bahasa</i> .....	IV-39
Gambar IV- 8 <i>Activity diagram upload dataset</i> .....	IV-40
Gambar IV- 9 <i>Activity diagram preprocessing dataset</i> .....	IV-41
Gambar IV- 10 <i>Activity diagram</i> melakukan proses <i>Word2vec</i> .....	IV-41
Gambar IV- 11 <i>Activity diagram training model LSTM</i> .....	IV-42
Gambar IV- 12 <i>Activity diagram</i> melakukan identifikasi bahasa .....	IV-42
Gambar IV- 13 <i>Sequence Diagram Upload Dataset</i> .....	IV-43
Gambar IV- 14 <i>Sequence Diagram Preprocessing Dataset</i> .....	IV-43
Gambar IV- 15 <i>Sequence Diagram Word2vec</i> .....	IV-44
Gambar IV- 16 <i>Sequence Diagram Training Model LSTM</i> .....	IV-44
Gambar IV- 17 <i>Sequence Diagram</i> Melakukan identifikasi bahasa .....	IV-44
Gambar IV- 18 <i>Class Diagram</i> .....	IV-45
Gambar IV- 19 Tampilan awal <i>Graphical user interface</i> (GUI) .....	IV-48
Gambar IV- 20 Tampilan <i>upload dataset Graphical user interface</i> .....	IV-48
Gambar IV- 21 Tampilan <i>preprocessing Graphical user interface</i> .....	IV-49
Gambar IV- 22 Tampilan <i>Word2vec Graphical user interface</i> .....	IV-49
Gambar IV- 23 Tampilan <i>Training model Graphical user interface</i> .....	IV-50

Gambar IV- 24 Tampilan Identifikasi bahasa <i>Graphical user interface</i> .....	IV-50
Gambar V- 1 Grafik Model <i>Accuracy</i> .....	V-6
Gambar V- 2 Grafik Model <i>Loss</i> .....	V-7

## DAFTAR ISTILAH

<i>Activity Diagram</i>	: Jenis diagram dalam digunakan untuk menggambarkan alur kerja atau proses dalam sistem
<i>Batch Size</i>	: Jumlah data yang diolah oleh komputer sekaligus saat belajar
<i>Black Box</i>	: Sistem atau objek di mana operasinya dimengerti dan diamati dari luar tanpa pengetahuan tentang proses internal atau mekanisme yang ada di dalamnya
<i>Class Diagram</i>	: Jenis diagram dalam yang digunakan untuk memodelkan struktur kelas dalam sebuah sistem
<i>Confusion Matrix</i>	: Alat yang digunakan untuk memahami kinerja model klasifikasi, yaitu model yang membedakan antara berbagai kelas atau kategori
<i>Dropout</i>	: Teknik regularisasi yang digunakan untuk mengurangi <i>overfitting</i> dalam proses pelatihan
<i>Encoding</i>	: Proses dalam penciptaan pesan melalui kode-kode tertentu agar dapat dibaca oleh sistem
<i>Epoch</i>	: Istilah yang digunakan untuk menggambarkan satu putaran lengkap di mana algoritma memproses seluruh data latih
<i>Field</i>	: Kolom didalam tabel yang menunjukkan suatu item data
<i>Flowchart</i>	: Gambaran ilustrasi langkah-langkah dan keputusan yang dilakukan dalam menjalankan suatu proses dari sebuah program
<i>F1-Score</i>	: Metrik yang menggabungkan precision dan recall menjadi satu ukuran kinerja tunggal
<i>Hidden Unit</i>	: Elemen dasar dalam lapisan tersembunyi dari jaringan saraf tiruan
<i>Hyperparameter</i>	: Parameter konfigurasi yang digunakan untuk mengatur proses pembelajaran pada algoritma pembelajaran mesin
<i>Input</i>	: Proses memasukkan data
<i>Layer</i>	: Lapisan suatu unit pada model
<i>Memory cell</i>	: Blok dasar dari memori komputer
<i>Multi-Class</i>	: Model klasifikasi yang memiliki setidaknya dua label berbeda
<i>Output</i>	: Hasil dari suatu proses pengolahan data <i>input</i>

<i>Precision</i>	: Ukuran yang mengevaluasi ketepatan model klasifikasi dalam mengidentifikasi kategori atau kelas yang benar
<i>Reccurent Dropout</i>	: Teknik regularisasi yang dirancang khusus untuk digunakan dalam jaringan saraf tiruan untuk mengurangi <i>overfitting</i>
<i>Recall</i>	: Ukuran yang mengevaluasi seberapa baik model klasifikasi mengidentifikasi semua kasus relevan dalam data
<i>Roman-aplhabet</i>	: Alfabet latin atau abjad romawi
<i>Sequence Diagram</i>	: jenis diagram yang digunakan dalam untuk menggambarkan interaksi antar objek dalam suatu sistem berdasarkan urutan waktu
<i>Sequential</i>	: Data yang berurutan
<i>Sigmoid</i>	: Fungsi aktivasi pada neuron
<i>Time step</i>	: Kejadian tunggal sel
<i>Use Case</i>	: Deskripsi terstruktur dari cara penggunaan sistem atau produk untuk mencapai tujuan tertentu
<i>User requirement</i>	: Kebutuhan pengguna yang disediakan oleh sistem
<i>Vanishing gradient</i>	: Masalah gradien hilang saat melatih jaringan saraf tiruan
<i>Word embedding</i>	: Proses konversi sebuah teks menjadi angka



# **BAB I**

## **PENDAHULUAN**

### **1.1 Pendahuluan**

Pada bab pendahuluan akan membahas latar belakang masalah, rumusan masalah, tujuan penelitian, manfaat penelitian, batasan masalah serta sistematika penulisan. Bab ini akan memberikan penjelasan umum mengenai keseluruhan isi pada Bab selanjutnya. Pendahuluan dimulai dengan penjelasan mengenai masalah yang ada dan bagaimana penyelesaian suatu masalah.

### **1.2 Latar Belakang**

Bahasa adalah alat komunikasi sosial yang terorganisasi dalam bentuk seperti kata, frasa, klausa dan kalimat yang digunakan baik secara lisan atau tulis. Bahasa berperan sangat penting bagi kehidupan manusia sebagai makhluk sosial, dengan adanya bahasa manusia dapat berinteraksi antara satu dan yang lainnya. Bahasa juga penting ketika kita akan mengembangkan empat keterampilan bahasa, yaitu berbicara, menyimak, membaca, dan menulis (Noermanzah, 2019).

Dengan adanya banyak bahasa yang digunakan di seluruh dunia, keahlian untuk mengidentifikasi bahasa menjadi semakin penting. Identifikasi bahasa adalah proses untuk menentukan bahasa yang digunakan dalam sebuah teks atau dokumen tertentu. Proses tersebut menggunakan teknologi pada bidang *Natural Language Processing* (NLP) yang memungkinkan sistem untuk menganalisis teks dan mengenali bahasa yang digunakan.

Untuk membuat sebuah sistem yang dapat mengidentifikasi bahasa, diperlukan sebuah metode yang mampu melakukan tugas tersebut. Salah satu metode yang dapat digunakan untuk mengidentifikasi bahasa adalah *Long Short-Term Memory* (LSTM). Metode ini merupakan pengembangan dari *Recurrent Neural Network* (RNN). Memori lama yang tersimpan pada RNN akan semakin tidak berguna dan tertimpa dengan memori baru. LSTM mampu mengatasi kendala tersebut karena dapat mengatur memori pada setiap masukannya dengan menggunakan *memory cells* dan *gate units* (Lubis & Kharisudin, 2021).

LSTM dapat mengingat informasi yang terkait dengan kata-kata pada teks yang telah diproses, penggunaan metode LSTM telah diterapkan dalam berbagai bidang, seperti pengenalan teks, klasifikasi dokumen, dan deteksi spam. Sebelumnya ada penelitian menggunakan metode serupa dengan judul Analisis Sentimen *Multi-Class* pada Sosial Media menggunakan metode *Long Short-Term Memory* (LSTM). Dengan total data ulasan novel yang digunakan adalah 400 data dibagi menjadi 8 kelas yaitu *Desire, acceptance, courage, peace, fear, pride, anger, love* dan mendapatkan hasil akurasi 89,45% dari 40 kali *epoch* (Astari et al., 2021).

Penelitian lainnya dengan pembahasan serupa namun metode yang berbeda dengan judul *Bhasha-Abhijnaanam: Native-script and Romanized Language Identification for 22 Indic Languages* dan mendapatkan akurasi sebesar 80,40% (Madhani et al., 2023). Berdasarkan penelitian yang sudah ada sebelumnya, identifikasi bahasa menggunakan metode LSTM merupakan topik penelitian yang menarik dalam bidang pengolahan bahasa alami.

Penelitian ini akan membangun sistem yang mampu mengidentifikasi bahasa menggunakan metode *Long Short Term Memory* (LSTM).

### **1.3 Rumusan Masalah**

Rumusan masalah dalam penelitian ini adalah sebagai berikut :

1. Bagaimana menerapkan metode *Long Short-Term Memory* (LSTM) untuk mengidentifikasi bahasa?
2. Bagaimana performa model LSTM dalam mengidentifikasi bahasa pada teks?

### **1.4 Tujuan Penelitian**

Tujuan penelitian ini adalah sebagai berikut :

1. Membangun sistem identifikasi bahasa menggunakan metode LSTM yang dapat digunakan dalam pengenalan bahasa.
2. Mengetahui akurasi metode LSTM dalam mengidentifikasi bahasa.

### **1.5 Manfaat Penelitian**

Manfaat Penelitian yang diperoleh adalah :

1. Membantu menyelesaikan permasalahan dalam memahami teks dengan bahasa yang berbeda-beda.
2. Membantu meningkatkan efisiensi pemrosesan teks yang memerlukan identifikasi bahasa.
3. Dapat digunakan sebagai rujukan dalam penelitian berikutnya.

### **1.6 Batasan Masalah**

Batasan masalah dalam penelitian ini adalah sebagai berikut :

1. Publikasi yang digunakan dalam penelitian hanya terdiri dari 10 bahasa

yang meliputi *Arabic, Chinese, Dutch, English, French, Indonesian, Japanese, Korean, Russian, Spanish*.

## **1.7 Sistematika Penulisan**

Adapun sistematika penulisan pada penelitian ini sebagai berikut :

### **BAB I. PENDAHULUAN**

Pada bab ini diuraikan mengenai latar belakang, perumusan masalah, tujuan penelitian, manfaat penelitian, batasan masalah/ruang lingkup, dan sistematika penulisan dalam penelitian ini.

### **BAB II. KAJIAN LITERATUR**

Pada bab ini akan dibahas dasar-dasar teori yang digunakan dalam penelitian, seperti definisi-definisi *Natural Language Processing, Language identification, Preprocessing Text, Word2vec, Artificial Neural Network, Long Short Term Memory, Confusion Matrix, Rational Unified Process* dan juga menguraikan penelitian-penelitian terdahulu yang relevan dengan penelitian ini.

### **BAB III. METODOLOGI PENELITIAN**

Pada bab ini akan membahas mengenai tahapan yang akan dilaksanakan pada penelitian ini mengenai metodologi dan tahapan perancangan penelitian seperti pengumpulan data, metode pengembangan perangkat lunak, dan manajemen proyek penelitian.

### **BAB IV. PENGEMBANGAN PERANGKAT LUNAK**

Pada bab ini akan membahas proses perancangan perangkat lunak yang akan dikembangkan. Dimulai dari analisis kebutuhan, perancangan,

konstruksi dan yang akhirnya akan dilakukan pengujian untuk memastikan perangkat lunak sudah sesuai dengan kebutuhan penelitian.

## **BAB V. HASIL DAN ANALISIS PENELITIAN**

Pada bab ini akan membahas hasil pengujian berdasarkan langkahlangkah yang telah ditetapkan sebelumnya. Tabel hasil pengujian akan menjadi acuan dalam membuat kesimpulan pada bab berikutnya.

## **BAB VI. KESIMPULAN DAN SARAN**

Pada bab ini akan membahas kesimpulan secara keseluruhan dari hasil penelitian yang telah dilakukan.

### **1.8 Kesimpulan**

Bab ini telah membahas mengenai latar belakang masalah, rumusan masalah, tujuan penelitian, manfaat penelitian, Batasan masalah, dan sistematika penulisan. Berdasarkan uraian di atas, dapat disimpulkan bahwa dalam penelitian ini dikembangkan sebuah metode yaitu *Long Short Term Memory* (LSTM) yang dapat mengidentifikasi bahasa pada pemrosesan teks.



## DAFTAR PUSTAKA

- Antares, J. 2020. Artificial Neural Network Dalam Mengidentifikasi Penyakit Stroke Menggunakan Metode Backpropagation (Studi Kasus di Klinik Apotik Madya Padang). *Djtechno: Journal of Information Technology Research*, Volume 1 Number 1, (<https://doi.org/10.46576/djtechno.v1i1.965>, diakses 16 Juli 2023) .
- Arjun M, & Shibu V. 2020. Malayalam Text AutoCorrection Using NLP. *Strad Research*, Volume 7 Number 8, (<https://doi.org/10.37896/sr7.8/049>, diakses 16 Juli 2023)
- Astari, Y., Afiyanti, & Rozaqi, S. W. 2021. Analisis Sentimen Multi-Class pada Sosial Media menggunakan metode Long Short-Term Memory (LSTM). *JLK*, 4(1), 8–12.
- Awaludin, M., & Raveena, R. R. 2021. Penerapan Metode Rational Unified Process Pada Knowledge Management System Untuk Mendukung Proses Pembelajaran Sekolah Menengah Atas. *JSI (Jurnal Sistem Informasi)*, (<https://doi.org/10.35968/jsi.v8i2.722>, diakses 22 Agustus 2023)
- Cahyudi, R. M., & Setiawan, E. B. 2023. Ekspansi Fitur dengan Word2vec dalam klasifikasi Hoax di Twitter. *E-Proceeding of Engineering*, 10(2), 1765–1776.
- Dewananda, K. F., Rahmawati, W. M., Wardhana, S. R., & Yuliasuti, G. E. 2022. Penentuan Relevansi Artikel Ilmiah dengan Metode Word2Vec. *Jurnal Riset Inovasi Bidang Informatika Dan Pendidikan Informatika (KERNEL)*, Volume 3 Number 2, (<https://doi.org/10.31284/j.kernel.2022.v3i2.4038>, diakses 10 Juli 2023)
- Grandini, M., Bagli, E., & Visani, G. 2020. Metrics for Multi-Class Classification: an Overview, (<http://arxiv.org/abs/2008.05756>, diakses 23 Agustus 2023)
- Hartanto, H., Liong, T. H., & Martina, I. 2013. Sistem Wawancara Virtual untuk Penerimaan Mahasiswa Jurusan Teknik Informatika di ITHB dengan Metode Natural Language Processing. *Jurnal Telematika*, Volume 8 Number 1, (<https://doi.org/10.61769/jurtel.v8i1.69>, diakses 16 Juli 2023)

- Jarvis, S., Bestgen, Y., & Pepper, S. 2013. Maximizing Classification Accuracy in Native Language Identification. *Association for Computational Linguistics*, 111–118.
- Le, X. H., Ho, H. V., Lee, G., & Jung, S. (2019). Application of Long Short-Term Memory (LSTM) neural network for flood forecasting. *Water (Switzerland)*, Volume 11 Number 7, (<https://doi.org/10.3390/w11071387>, diakses 16 Juli 2023)
- Lotfi, E., Markov, I., & Daelemans, W. 2020. A Deep Generative Approach to Native Language Identification. *Proceedings of the 28th International Conference on Computational Linguistics*, 1778–1783.
- Lubis, J. K., & Kharisudin, I. 2021. Metode Long Short Term Memory dan Generalized Autoregressive Conditional Heteroscedasticity untuk Pemodelan Data Saham. *PRISMA, Prosiding Seminar Nasional Matematika*, Volume 4, (<https://journal.unnes.ac.id/sju/index.php/prisma/>, diakses 16 Juli 2023)
- Madhani, Y., Khapra, M. M., & Kunchukuttan, A. 2023. Bhasha-Abhijnaanam: Native-script and romanized Language Identification for 22 Indic languages. (<http://arxiv.org/abs/2305.15814>, 12 Agustus 2023)
- Malmasi, S. 2016. Native Language Identification: Explorations and Applications. Tesis Macquarie University.
- Manaswi, N. K. 2018. Deep Learning with Applications Using Python. In *Deep Learning with Applications Using Python*. Bangalore. (<https://doi.org/10.1007/978-1-4842-3516-4>, 16 Juli 2023)
- McNamee, P. 2005. Language Identification: A Solved Problem Suitable for Undergraduate Instruction. *Journal of Computing Sciences in Colleges*, Volume 20 Number 3 (<https://dl.acm.org/doi/10.5555/1040196.1040208>, diakses 9 Juli 2023)
- Nawangsih, I., Melani, I., & Fauziah, S. 2021. Prediksi Pengangkatan Karyawan Dengan Metode Algoritma C5.0. *Jurnal Pelita Teknologi*, Volume 16 Number 2, (<https://doi.org/10.37366/pelitatekno.v16i2.672>, diakses 30 Juli 2023)

- Noermanzah. 2019. Bahasa sebagai Alat Komunikasi, Citra Pikiran dan Kepribadian. *Semiba*, 306–319.
- Ramadhanti, N. R., & Mariyah, S. 2019. Document Similarity Detection Using Indonesian Language Word2vec Model. 2019 3rd International Conference on Informatics and Computational Sciences, 1–6.
- Rowan, Muflikhah, L., & Cholissodin, I. 2022. Peramalan Kasus Positif COVID-19 di Jawa Timur menggunakan Metode Hybrid ARIMA-LSTM. *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer*, Volume 6 Number 9, (<http://j-ptiik.ub.ac.id>, diakses 22 Agustus 2023)
- Santoso, J., Setiawan, E. I., Purwanto, C. N., & Kurniawan, F. 2021. Indonesian Sentence Boundary Detection using Deep Learning Approaches. *Knowledge Engineering and Data Science*, Volume 4 Number 1, (<https://doi.org/10.17977/um018v4i12021p38-48>, diakses 16 Juli 2023)
- Saputro, I. W., & Sari, B. W. 2019. Uji Performa Algoritma Naïve Bayes untuk Prediksi Masa Studi Mahasiswa Naïve Bayes Algorithm Performance Test for Student Study Prediction. *Citec Journal*, Volume 6 Number 1, (<https://doi.org/10.24076/citec.2019v6i1.178>, diakses 30 Juli 2023)
- Wicaksono, A. A., Yusuf, R., & Saputri, T. A. 2021. Penerapan Natural Language Processing Berbasis Virtual Assistant Pada Bagian Administrasai Akademik STMIK Dharma Wacana. *Jurnal IRobot*, Volume 5, (<https://doi.org/10.53514/ir.v5i1.228>, diakses 16 Juli 2023)
- Wijaya, A. H. 2019. Artificial Neural Network Untuk Memprediksi Beban Listrik Dengan Menggunakan Metode Backpropagation. *Jurnal CoreIT*, Volume 5 Number 2, (<http://dx.doi.org/10.24014/coreit.v5i2.8280>, diakses 16 Juli 2023)
- Windarto, A. P., Lubis, M. R., & Solikhun. 2018. Model Arsitektur Neural Network Dengan Backpropogation Pada Prediksi Total Laba Rugi Komprehensif Bank Umum Konvensional. *Kumpulan Jurnal Ilmu Komputer (KLIK)*, Volume 5 Number 2, (<http://dx.doi.org/10.20527/klik.v5i2>, diakses 17 Juli 2023)

- Yunefri, Y., Fadrial, Y. E., & Sutejo. 2021. Chatbot Pada Smart Cooperative Oriented Problem Menggunakan Natural Language Processing dan Naive Bayes Classifier. *Journal of Information Technology and Computer Science (INTECOMS)*, Volume 4 Number 2, (<https://doi.org/10.31539/intecom.v4i2.2704>, diakses 9 Agustus 2023)
- Zampieri, M., Ciobanu, A. M., & Dinu, L. P. 2017. Native Language Identification on Text and Speech. (<http://arxiv.org/abs/1707.07182>, diakses 18 Juli 2023)