

NATURAL LANGUAGE PROCESSING

Dasar dan Konsep Bagi Pemula

JILID 1



Rizal Rachman, S.Si., M.M., M.Kom | Syarif Hidayatulloh, S.T., M.Kom
Salman Topiq, M.Kom | Rangga Sanjaya, S.T., M.Kom

NATURAL LANGUAGE PROCESSING

Dasar dan
Konsep Bagi
Pemula

JILID 1

Rizal Rachman., S.Si., M.M., M.Kom.

Syarif Hidayatulloh, S.T., M.Kom.

Salman Topiq, M.Kom.

Rangga Sanjaya, S.T., M.Kom.



PENERBIT KBM INDONESIA

Adalah penerbit dengan misi memudahkan proses penerbitan buku
buku penulis di tanah air Indonesia. Serta menjadi media sharing
proses penerbitan buku.

NATURAL LANGUAGE PROCESSING:

Dasar dan Konsep Bagi Pemula Jilid 1

Copyright @2025 By Rizal Rachman., S.Si., M.M., M.Kom., dkk

All right reserved

Penulis

Rizal Rachman., S.Si., M.M., M.Kom.

Syarif Hidayatulloh, S.T., M.Kom.

Salman Topiq, M.Kom.

Rangga Sanjaya, S.T., M.Kom.

Desain Sampul

Aswan Kreatif

Tata Letak

Sofitahm

Editor

Dr. Muhamad Husein Maruahey, Drs., M.Sc.

Background isi buku di ambil dari <https://www.freepik.com/>

Official

Depok, Sleman-Jogjakarta (Kantor)

Penerbit Karya Bakti Makmur (KBM) Indonesia

Anggota IKAPI/No. IKAPI 279/JTI/2021

081357517526 (Tlpn/WA)

Website

<https://penerbitkbm.com>

www.penerbitbukumurah.com

Email

naskah@penerbitkbm.com

Distributor

<https://penerbitkbm.com/toko-buku/>

Youtube

Penerbit KBM Sastrabook

Instagram

[@penerbit.kbmindonesia](https://www.instagram.com/penerbit.kbmindonesia)

[@penerbitbukujogja](https://www.instagram.com/penerbitbukujogja)

ISBN (Lengkap): 978-634-202-843-8

ISBN (Jilid 1): 978-634-202-844-5

Cetakan ke-1, Desember 2025

15,5 x 23 cm, vi + 305 halaman

Isi buku diluar tanggungjawab penerbit
Hak cipta merek KBM Indonesia sudah terdaftar di DJKI-Kemenkumham
dan isi buku dilindungi undang-undang.
Dilarang keras menerjemahkan, memfotokopi, atau
memperbanyak sebagian atau seluruh isi buku ini
tanpa seizin penerbit karena beresiko sengketa hukum

Sanksi Pelanggaran Pasal 113
Undang-Undang No. 28 Tahun 2014 Tentang Hak Cipta

1. Setiap Orang yang dengan tanpa hak melakukan pelanggaran hak ekonomi sebagaimana dimaksud dalam Pasal 9 ayat (1) huruf i untuk Penggunaan Secara Komersial dipidana dengan pidana penjara paling lama 1 (satu) tahun dan/atau pidana denda paling banyak Rp 100.000.000 (seratus juta rupiah).
2. Setiap Orang yang dengan tanpa hak dan/atau tanpa izin Pencipta atau pemegang Hak Cipta melakukan pelanggaran hak ekonomi Pencipta sebagaimana dimaksud dalam Pasal 9 ayat (1) huruf c, huruf d, huruf f, dan/atau huruf h untuk Penggunaan Secara Komersial dipidana dengan pidana penjara paling lama 3 (tiga) tahun dan/atau pidana denda paling banyak Rp 500.000.000,00 (lima ratus juta rupiah).
3. Setiap Orang yang dengan tanpa hak dan/atau tanpa izin Pencipta atau pemegang Hak Cipta melakukan pelanggaran hak ekonomi Pencipta sebagaimana dimaksud dalam Pasal 9 ayat (1) huruf a, huruf b, huruf e, dan/atau huruf g untuk Penggunaan Secara Komersial dipidana dengan pidana penjara paling lama 4 (empat) tahun dan/atau pidana denda paling banyak Rp 1.000.000.000,00 (satu miliar rupiah).
4. Setiap Orang yang memenuhi unsur sebagaimana dimaksud pada ayat (3) yang dilakukan dalam bentuk pembajakan, dipidana dengan pidana penjara paling lama 10 (sepuluh) tahun dan/atau pidana denda paling banyak Rp 4.000.000.000,00 (empat miliar rupiah).

KATA PENGANTAR

Segala puja dan puji syukur kehadiran Allah Swt., yang telah melimpahkan rahmat dan hidayah-Nya kepada kita semua, Shalawat dan salam tetap tercurah pada junjungan Nabi Kita Muhammad SAW. Alhamdulillah atas Rahmat dan Karunia-Nyalah kami dapat menyelesaikan buku dengan judul "**Natural Language Processing: Dasar dan Konsep Bagi Pemula**". Buku ini disusun sebagai upaya untuk memberikan pemahaman mendalam mengenai Natural Language Processing (NLP), salah satu cabang dari kecerdasan buatan (artificial intelligence) yang berkembang pesat di dunia teknologi informasi.

Natural Language Processing (NLP) adalah bidang yang menggabungkan ilmu komputer, linguistik, dan statistik untuk memungkinkan komputer memahami, menganalisis, dan menghasilkan bahasa manusia. Seiring dengan pesatnya perkembangan teknologi dan penggunaan data besar (big data), aplikasi NLP kini telah merambah banyak sektor, mulai dari asisten virtual, analisis sentimen, hingga penerjemahan otomatis. Buku ini hadir untuk memperkenalkan konsep dasar, teknik-teknik utama, serta berbagai aplikasi NLP dalam kehidupan sehari-hari.

Penyusunan buku ini tidak lepas dari dukungan berbagai pihak yang telah memberikan kontribusi berharga, baik berupa ide, referensi, maupun masukan yang sangat bermanfaat. Kami juga berharap buku ini dapat menjadi sumber referensi yang berguna bagi para mahasiswa, peneliti, maupun praktisi yang tertarik untuk mendalami dan mengembangkan teknologi NLP. Dengan begitu,

kami berharap buku ini dapat memberikan wawasan dan pemahaman yang lebih baik mengenai bagaimana teknologi NLP bekerja dan potensinya untuk masa depan.

Akhir kata, kami menyadari bahwa buku ini masih jauh dari sempurna. Oleh karena itu, kritik dan saran yang membangun akan sangat kami terima demi perbaikan di masa mendatang. Semoga buku ini dapat memberikan manfaat yang besar bagi pembaca dan mendorong perkembangan penelitian serta penerapan NLP di Indonesia dan dunia.

DAFTAR ISI

KATA PENGANTAR	i
DAFTAR ISI	iii
BAB 1 PENDAHULUAN NATURAL LANGUAGE PROCESSING (NLP)	1
1.1 Definisi dan Pentingnya NLP	1
1.2 Sejarah Singkat NLP	3
1.3 Aplikasi NLP di Dunia Nyata	5
1.4 Tantangan dalam NLP	9
1.5 Perspektif Interdisipliner dalam NLP (Komputasi, Linguistik, dan Statistik)	13
BAB 2 DASAR-DASAR LINGUISTIK UNTUK NLP	19
2.1 Pengantar Linguistik	19
2.2 Fonetik dan Fonologi: Dasar Pengolahan Suara	19
2.3 Morfologi: Struktur Kata dan Bentuknya	20
2.4 Sintaksis: Struktur Kalimat dan Penguraiannya	20
2.5 Semantik: Pemahaman Makna dalam Bahasa	21
2.6 Pragmatik: Konteks dan Makna yang Terkait	21
2.7 Sociolinguistik dan Variasi Bahasa	21
2.8 Pendekatan Statistik dan Pembelajaran Mesin	22
2.9 Corpus Linguistics	22
BAB 3 REPRESENTASI TEKS DALAM NLP	23
3.1 Pengenalan Representasi Teks	23
3.2 Model Representasi Kata: BOW, TF-IDF, dan Word Embeddings	23
3.3 Word2Vec dan GloVe	28
3.4 Representasi Kalimat: Sentence Embeddings	32

3.5	Pretrained Language Models: BERT, GPT, dan Transformernya	38
BAB 4 PREPROCESSING DATA TEKS		45
4.1	Pengantar Preprocessing dalam NLP	45
4.2	Tokenisasi: Mengubah Teks Menjadi Unit yang Dapat Diproses	45
4.3	Penghilangan Stop Words, Stemming, dan Lematisasi	49
4.4	Normalisasi Teks: Lowercasing, Punctuation Removal, dll.	53
4.5	Bagaimana Memilih Teknik Preprocessing yang Tepat?	58
BAB 5 PENGELOLAAN DATA TEKS DALAM NLP		63
5.1	Pengumpulan dan Pembersihan Data Teks	63
5.2	Teknik Sampling Data dalam NLP	69
5.3	Pengolahan Data Tidak Terstruktur	75
5.4	Mengatasi Masalah Ketidakseimbangan Data dalam NLP	81
5.5	Data Augmentation dalam NLP	87
BAB 6 PEMBELAJARAN MESIN DALAM NLP		95
6.1	Pengantar Pembelajaran Mesin untuk NLP	95
6.2	Model Klasifikasi: Naive Bayes, SVM, dan KNN	96
6.3	Model Regresi: Linear dan Logistic Regression	102
6.4	Pembelajaran Terawasi dan Tak Terawasi dalam NLP	106
6.5	Evaluasi Model: Akurasi, Presisi, Recall, dan F1 Score	112
BAB 7 DEEP LEARNING DAN NEURAL NETWORKS DALAM NLP		119
7.1	Apa Itu Deep Learning?	119
7.2	Neural Networks untuk NLP	120
7.3	RNN dan LSTM: Model untuk Pengolahan Urutan	123
7.4	CNN untuk Ekstraksi Fitur Teks	127
7.5	Transformer: Dasar dan Penerapan dalam NLP	131

BAB 8 MODEL BAHASA DAN PEMBELAJARAN REPRESENTASI -----	137
8.1 Konsep Model Bahasa dalam NLP -----	137
8.2 Pengenalan Pembelajaran Representasi Kata -----	138
8.3 Pretrained Models: BERT, GPT, dan Variannya -----	139
8.4 Fine-Tuning Pretrained Models untuk Tugas Tertentu -----	144
8.5 Transfer Learning dalam NLP -----	149
BAB 9 PEMROSESAN TEKS UNTUK TUGAS-TUGAS NLP ----	155
9.1 Named Entity Recognition (NER) -----	155
9.2 Part-of-Speech Tagging -----	160
9.3 Dependency Parsing -----	165
9.4 Sentiment Analysis dan Opinion Mining -----	171
9.5 Text Summarization dan Generation -----	177
BAB 10 ANALISIS TEKS DAN PENCARIAN INFORMASI-----	183
10.1 Pengenalan Pencarian Informasi (IR) -----	183
10.2 Teknik Pencarian Berbasis Teks: TF-IDF dan Cosine Similarity-----	185
10.3 Sistem Rekomendasi Teks -----	189
10.4 Analisis Sentimen dan Evaluasi Teks-----	195
10.5 Pencarian dan Pengambilan Informasi pada Dokumen Besar-----	200
BAB 11 PEMROSESAN BAHASA UNTUK PENGENALAN SUARA DAN SPEECH-TO-TEXT -----	207
11.1 Teknologi Pengenalan Suara-----	207
11.2 Speech-to-Text: Sistem dan Algoritma Dasar -----	211
11.3 Tantangan dalam Pengolahan Ucapan -----	215
11.4 Aplikasi NLP pada Pengenalan Suara-----	219
11.5 Pengolahan Bahasa dalam Sistem Voice Assistants---	223
BAB 12 EVALUASI DAN UKURAN KINERJA DALAM NLP----	229
12.1 Pentingnya Evaluasi dalam NLP-----	229
12.2 Metode Evaluasi Model NLP: Cross-Validation, Hold-out, dll. -----	233
12.3 Metrik Evaluasi: Akurasi, Perplexity, BLEU, ROUGE ----	238

12.4 Evaluasi dalam Tugas-Tugas Tertentu: Klasifikasi, Generasi Teks, dll. -----	243
12.5 Masalah Umum dalam Evaluasi Model NLP -----	249
BAB 13 TANTANGAN DAN PERKEMBANGAN MASA DEPAN NLP -----	255
13.1 Tantangan dalam Pemrosesan Bahasa Alami-----	255
13.2 NLP Multibahasa dan Permasalahan dalam Penanganan Bahasa Daerah -----	260
13.3 NLP dan Kecerdasan Buatan Umum (AGI)-----	265
13.4 Etika dalam NLP: Bias, Privasi, dan Penggunaan yang Adil -----	270
13.5 Arah Penelitian dan Perkembangan NLP di Masa Depan -----	275
BAB 14 APLIKASI NLP DALAM BERBAGAI BIDANG-----	283
14.1 NLP dalam Bidang Kesehatan: Diagnosa dan Analisis Data Medis -----	283
14.2 NLP untuk Pengolahan Data Bisnis: Chatbots dan Virtual Assistants-----	289
14.3 NLP dalam Media Sosial: Analisis Sentimen dan Pemetaan Emosi-----	294
DAFTAR PUSTAKA -----	301
PROFIL PENULIS-----	303

DAFTAR PUSTAKA

- Bengio, Y., Courville, A., & Vincent, P. (2013). *Learning Deep Architectures for AI*. Foundations and Trends® in Machine Learning.
- Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media.
- Chomsky, N. (2006). *The Minimalist Program*. MIT Press.
- Goldberg, Y. (2017). *Neural Network Methods for Natural Language Processing*. Morgan & Claypool Publishers.
- Goldberg, Y. (2020). *Transformers for Natural Language Processing: Build state-of-the-art natural language processing systems using the transformer architecture*. Packt Publishing.
- Hirschberg, J., & Manning, C. D. (2015). *Advances in Natural Language Processing*. Cambridge University Press.
- Jurafsky, D., & Martin, J. H. (2008). *Speech and Language Processing (2nd ed.)*. Pearson.
- Jurafsky, D., & Martin, J. H. (2023). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition (3rd ed.)*. Pearson.
- Koehn, P. (2009). *Statistical Machine Translation*. Cambridge University Press.
- Liddy, E. D. (2001). *Natural Language Processing and Information Retrieval*. Kluwer Academic Publishers.
- Lin, C. Y. (2006). *Automatic Text Summarization*. In *Handbook of Natural Language Processing (2nd ed.)*, CRC Press.

- Manning, C. D., & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press.
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). *Distributed Representations of Words and Phrases and Their Compositionality*. Advances in Neural Information Processing Systems (NeurIPS).
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, Ł., & Polosukhin, I. (2017). *Attention is All You Need*. Advances in Neural Information Processing Systems (NeurIPS).
- Yang, D., & Salakhutdinov, R. (2018). *Deep Learning for NLP and Speech Recognition*. Cambridge University Press.

PROFIL PENULIS

■ BIODATA PENULIS 1



Rizal Rachman., S.Si., M.M., M.Kom, lahir di Indramayu, Jawa Barat pada tanggal 10 November 1982. Saat ini sedang menempuh studi Doktor Sistem Informasi di UNIKOM Bandung (2025), Magister Sistem Informasi di STMIK LIKMI Bandung (2021), Magister Manajemen SDM di Universitas BSI Bandung (2014), Sarjana Ilmu Komputer di UNPAD Bandung (2005). Penulis adalah dosen tetap

dan pengajar di Fakultas Teknologi Informasi, Program Studi Sistem informasi, Universitas Ardhirajasa Reswara Sanjaya. Selain itu penulis sebagai Direktur Sekolah Non Formal di Homeschooling Milenial Unggul Bandung dan Koordinator Tutor PKBM Minda Utama Bandung. Penulis menghasilkan lebih dari **84 publikasi ilmiah**, termasuk 5 **Jurnal Scopus** dan **79 artikel jurnal**, dengan **H-indeks 16** dan Sinta Score Overall 778. Minat penelitiannya mencakup **data Scientist, kecerdasan buatan, dan rekayasa perangkat lunak**. **Moto Hidup: Hidup Itu harus bisa bermanfaat untuk orang lain.**

■ BIODATA PENULIS 2



Syarif Hidayatulloh, S.T., M.Kom., lahir di Ciamis pada tanggal 27 Februari 1990. Saat ini sedang menempuh studi Doktoral (Ph.D) di bidang Information and Communication Technology (ICT) dengan fokus pada keamanan siber di perguruan tinggi Indonesia. Telah menyelesaikan program Magister Ilmu Komputer pada tahun 2015 di Universitas Nusa Mandiri,

Jakarta. Penulis merupakan dosen tetap di ARS University, Bandung. Selain aktif mengajar, penulis juga melakukan penelitian di bidang keamanan siber, tata kelola teknologi informasi, serta pengembangan kerangka kerja strategis untuk institusi pendidikan. Sejumlah karya ilmiah telah dipublikasikan di jurnal nasional maupun internasional, serta dipresentasikan dalam berbagai forum akademik.

■ BIODATA PENULIS 3



Salman Topiq, M.Kom., lahir di Ciamis pada tanggal 01 April 1993. Telah menyelesaikan program Magister Ilmu Komputer pada tahun 2018 di Universitas Nusa Mandiri. Penulis merupakan dosen tetap di Program Studi Teknik Informatika, Fakultas Teknologi dan Informasi, Universitas Adhirajasa Reswara Sanjaya, Bandung. Selain aktif menjadi dosen, beberapa

kegiatan baik internal maupun eksternal pernah diikuti; salah satunya sebagai instruktur pelatihan drone di Drone Edutech sejak tahun 2021 Hingga Sekarang

■ BIODATA PENULIS 4



Rangga Sanjaya, S.T., M.Kom., lahir di Bandung pada tanggal 2 Oktober 1988. Saat ini sedang menempuh studi Doktorat (Ph.D. Prestasi saya adalah terdaftarnya 13 HKI di Direktorat Jenderal Kekayaan Intelektual Republik Indonesia. Untuk pendidikan formal, saya lulusan STMIK Nusamandiri, Jakarta (M.Kom Ilmu Komputer). Minat saya adalah inovasi digital dan transformasi digital, dan salah satu topik perkuliahan saya adalah Ekosistem Inovasi Digital. Saat ini saya sedang melakukan penelitian di bidang transformasi digital dan sistem pendukung keputusan.