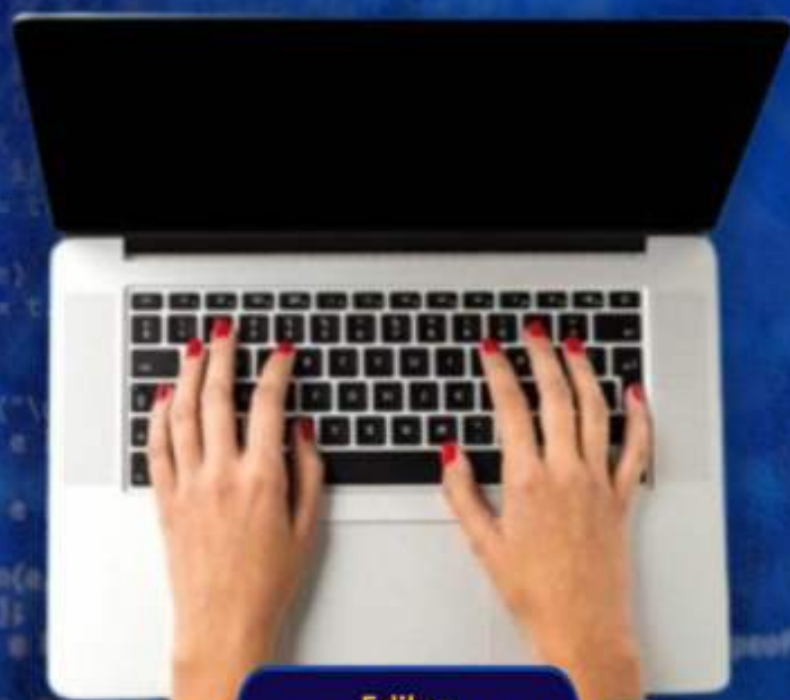


Alfiansyah Hasibuan, S.Kom, M.Kom
Abraham Yano Suharmanto, S.Kom., M.Kom
Prof. Dr. Ing. Drs. Parabelem Tinno Dolf Rompas, MT
Sapriadi, S.Kom, M.Kom



TEXT MINING *dan* NATURAL LANGUAGE PROCESSING



Editor:
Mayang Wulan, SST, M.K.M, P.hD.
Merriam Modeong, S.Pd, M.Pd.

TEXT MINING DAN NATURAL LANGUAGE PROCESSING

Alfiansyah Hasibuan, S.Kom, M.Kom

Abraham Yano Suharmanto, S.Kom., M.Kom

Prof. Dr. Ing. Drs. Parabelem Tinno Dolf Rompas, MT

Sapriadi, S.Kom, M.Kom



Tahta Media Group

UU No 28 tahun 2014 tentang Hak Cipta

Fungsi dan sifat hak cipta Pasal 4

Hak Cipta sebagaimana dimaksud dalam Pasal 3 huruf a merupakan hak eksklusif yang terdiri atas hak moral dan hak ekonomi.

Pembatasan Pelindungan Pasal 26

Ketentuan sebagaimana dimaksud dalam Pasal 23, Pasal 24, dan Pasal 25 tidak berlaku terhadap:

- penggunaan kutipan singkat Ciptaan dan/atau produk Hak Terkait untuk pelaporan peristiwa aktual yang ditujukan hanya untuk keperluan penyediaan informasi aktual;
- Penggandaan Ciptaan dan/atau produk Hak Terkait hanya untuk kepentingan penelitian ilmu pengetahuan;
- Penggandaan Ciptaan dan/atau produk Hak Terkait hanya untuk keperluan pengajaran, kecuali pertunjukan dan Fonogram yang telah dilakukan Pengumuman sebagai bahan ajar; dan
- penggunaan untuk kepentingan pendidikan dan pengembangan ilmu pengetahuan yang memungkinkan suatu Ciptaan dan/atau produk Hak Terkait dapat digunakan tanpa izin Pelaku Pertunjukan, Produser Fonogram, atau Lembaga Penyiaran.

Sanksi Pelanggaran Pasal 113

- 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 Rp100.000.000 (seratus juta rupiah).
- 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 Rp500.000.000,00 (lima ratus juta rupiah).

Hak cipta dilindungi undang-undang

Dilarang keras mereproduksi, merefotokopi, atau memperbanyak sebagian atau seluruh isi buku ini tanpa izin tertulis dari Penerbit.

TEXT MINING DAN NATURAL LANGUAGE PROCESSING

Penulis:

Alfiansyah Hasibuan, S.Kom, M.Kom
Abraham Yano Suharmanto, S.Kom., M.Kom
Prof. Dr. Ing. Drs. Parabelem Tinno Dolf Rompas, MT
Sapriadi, S.Kom, M.Kom

Desain Cover:

Tahta Media

Editor:

Mayang Wulan, SST, M.K.M, P.hD
Merriam Modeong, S.Pd, M.Pd

Proofreader:

Tahta Media

Ukuran:

x, 117, Uk: 15,5 x 23 cm

ISBN: 978-634-262-319-0

Cetakan Pertama:

Oktober 2026

Hak Cipta 2026, Pada Penulis

Isi diluar tanggung jawab penerbit

Copyright © 2026 by Tahta Media Group

All Right Reserved

PENERBIT TAHTA MEDIA GROUP

Perumahan Mitra Utama Residence 3 Blok A no 1, Sawahan, Tempel,
Kec. Gatak, Kabupaten Sukoharjo, Jawa Tengah 57557

email: tahtaliterasimedia@gmail.com

website: <http://store.tahtamedia.co.id/>

Anggota IKAPI (216/JTE/2021)

KATA PENGANTAR

Data tekstual merupakan salah satu bentuk data yang paling banyak dihasilkan manusia. Ulasan pelanggan, dokumen organisasi, berita, percakapan layanan, media sosial, artikel ilmiah, laporan publik, dan arsip digital menyimpan informasi yang bernilai, tetapi informasi tersebut tidak langsung tersedia dalam bentuk numerik yang mudah dianalisis. *Text mining* dan *natural language processing* menyediakan seperangkat konsep dan metode untuk mengubah bahasa manusia menjadi representasi yang dapat diproses komputer tanpa mengabaikan struktur, konteks, dan makna yang melekat pada bahasa.

Perkembangan NLP bergerak dari pendekatan berbasis aturan dan fitur manual menuju machine learning, representasi distribusional, *deep learning*, arsitektur Transformer, model prelatih, dan large language models. Perubahan tersebut meningkatkan kemampuan sistem dalam menangani konteks dan ragam tugas, tetapi sekaligus memperbesar kebutuhan akan evaluasi yang ketat, tata kelola data, mitigasi bias, perlindungan privasi, serta pemahaman terhadap keterbatasan model generatif.

Buku ini menyusun perjalanan belajar secara berlapis. Pembaca mulai dari konsep teks dan korpus, fondasi linguistik, *preprocessing*, representasi fitur, eksplorasi data, klasifikasi, *sentiment analysis*, *topic modeling*, *information extraction*, *similarity* dan *information retrieval*.

Semoga buku ini membantu mahasiswa dan praktisi memahami text mining dan NLP sebagai disiplin yang menghubungkan ilmu komputer, statistika, linguistik, machine learning, dan rekayasa sistem. Kemampuan menggunakan model modern penting, tetapi kemampuan merumuskan masalah dan mengevaluasi bukti tetap menjadi inti pekerjaan ilmiah.

PRAKATA

Alhamdulillahirabbil'alamiin, segala puja dan puji syukur penulis panjatkan kepada Allah Yang Maha Pengasih dan Maha Penyayang, yang kasih dan sayangnya tiada batas kepada seluruh Makhluk ciptan-Nya. Tanpa karunia-Nya tidak mungkin naskah buku ajar yang berjudul *Text Mining Dan Natural Language Processing* dapat terselesaikan dengan tepat waktu mengingat tugas dan kewajiban lain yang bersamaan hadir. Dengan terselesaikannya penulisan buku ajar ini juga tidak terlepas dari bantuan berbagai pihak. Buku ajar *Text Mining Dan Natural Language Processing* ini bertujuan untuk memberikan pengetahuan kepada pembaca terutama kepada mahasiswa, karena buku ajar ini penulis desain menyesuaikan dengan rencana Pembelajaran Semester (RPS). Buku ajar ini menjadi acuan penulis dalam membuat materi perkuliahan.

Saya ucapkan terimakasih yang tiada tara kepada kedua orang tua, kepada istri dan anak yang selalu memberi support kepada penulis tanpa kenal lelah. Penulis ucapkan terimakasih kepada pihak Fakultas Teknik. Kepada Dekan Fakultas Teknik Universitas Negeri Manado saya ucapkan ribuan terimakasih, tidak lupa juga saya ucapkan terimakasih kepada pimpinan program studi Teknik Informatika, Kepada civitas akademika program studi Teknik Informatika saya ucapkan terimakasih. Kepada tim *teaching* mata kuliah *Text Mining Dan Natural Language Processing* terimakasih untuk semua bantuan, motivasi dan saran-sarannya. Meskipun telah berusaha untuk menghindari kesalahan, penulis menyadari juga bahwa buku ini masih mempunyai kelemahan dan kekurangan. Karena itu, penulis berharap agar pembaca berkenan menyampaikan kritikan. Dengan segala pengharapan dan

keterbukaan, penulis menyampaikan rasa terimakasih dengan setulus-tulusnya.

Semoga buku ini membantu mahasiswa dan praktisi memahami text mining dan NLP sebagai disiplin yang menghubungkan ilmu komputer, statistika, linguistik, *machine learning*, dan rekayasa sistem. Kemampuan menggunakan model modern penting, tetapi kemampuan merumuskan masalah dan mengevaluasi bukti tetap menjadi inti pekerjaan ilmiah.

Tondano, Agustus 2026

Penulis

DAFTAR ISI

KATA PENGANTAR.....	iv
PRAKATA	v
DAFTAR ISI.....	vii
BAB 1 FONDASI <i>TEXT MINING</i> DAN <i>NATURAL LANGUAGE</i> <i>PROCESSING</i>	1
A. Capaian Pembelajaran	1
B. Mengapa Data Teks Sulit Dianalisis	1
C. Text Mining dan NLP dalam Satu Ekosistem	5
D. Pipeline Proyek Teks.....	10
E. Pendekatan Rule-based, Statistik, dan Neural	12
F. Contoh Masalah.....	16
G. Latihan dan Diskusi	18
BAB 2 KORPUS, AKUISISI DATA, ANOTASI, DAN TATA KELOLA.....	19
A. Capaian pembelajaran	19
B. Korpus dan Representativitas.....	19
C. Akuisisi Data	22
D. Anotasi dan Skema Label.....	28
E. Latihan dan Diskusi	33
BAB 3 <i>PREPROCESSING</i> DAN NORMALISASI TEKS.....	34
A. Capaian Pembelajaran	34
B. Audit Sebelum <i>Cleaning</i>	34
C. Tokenization.....	39
D. <i>Case Folding</i> , <i>Stopword</i> , <i>Stemming</i> , dan <i>Lemma</i>	44
E. Normalisasi Bahasa Informal.....	49
F. Negasi, Emoji, URL, dan Entitas Khusus	55
G. Latihan dan Diskusi	58

BAB 4 REPRESENTASI TEKS DARI BAG OF WORDS SAMPAI EMBEDDING	59
A. Capaian Pembelajaran	59
B. <i>Bag of Words</i> dan <i>Document-Term Matrix</i>	59
C. N-gram	63
D. TF-IDF	69
E. <i>Word Embedding</i>	74
F. Latihan dan Diskusi	80
BAB 5 KLASIFIKASI TEKS DENGAN MACHINE LEARNING	81
A. Capaian Pembelajaran	81
B. Formulasi Klasifikasi	81
C. <i>Naive Bayes</i>	85
D. <i>Logistic Regression</i>	89
E. <i>Class Imbalance</i> dan <i>Threshold</i>	93
F. Latihan dan Diskusi	97
BAB 6 SENTIMENT ANALYSIS, EMOTION, DAN OPINION MINING	98
A. Capaian Pembelajaran	98
B. Definisi Sentimen dan Target	98
C. <i>Lexicon-based</i>	101
D. <i>Supervised Sentiment</i>	104
E. Latihan dan Diskusi	109
DAFTAR PUSTAKA	112
GLOSARIUM	114
PROFIL PENULIS	116

BAB 1

FONDASI *TEXT MINING* DAN *NATURAL LANGUAGE* *PROCESSING*

A. CAPAIAN PEMBEJALARAN

Mahasiswa dapat mengetahui dan memahami tentang :

1. Membedakan text mining, NLP, information retrieval, dan computational linguistics.
2. Menjelaskan pipeline umum proyek analisis teks.
3. Mengidentifikasi karakteristik data teks dan konsekuensinya terhadap pemodelan.

B. MENGAPA DATA TEKS SULIT DIANALISIS

Data teks merupakan salah satu bentuk data yang paling banyak dihasilkan dalam era digital. Berbagai aktivitas manusia seperti komunikasi melalui media sosial, transaksi digital, penulisan dokumen, ulasan pelanggan, artikel berita, laporan perusahaan, hingga rekam medis menghasilkan jumlah data teks yang terus meningkat. Berbeda dengan data numerik yang memiliki struktur jelas dan mudah diproses oleh komputer, data teks termasuk dalam kategori unstructured data atau data tidak terstruktur. Kondisi tersebut menyebabkan proses analisis teks menjadi jauh lebih kompleks karena komputer tidak secara langsung memahami bahasa manusia sebagaimana manusia memahaminya.

Kesulitan utama dalam menganalisis data teks berasal dari karakteristik bahasa manusia yang bersifat fleksibel, dinamis, dan sangat bergantung pada konteks. Manusia dapat memahami maksud sebuah kalimat melalui pengalaman, budaya, pengetahuan sebelumnya, serta situasi komunikasi. Sebaliknya, komputer hanya menerima data dalam bentuk simbol atau karakter yang harus diubah terlebih dahulu menjadi representasi matematis agar dapat diproses menggunakan algoritma.

Sebagai contoh, kalimat:

"Pelayanan restoran ini luar biasa, sampai saya tidak ingin kembali lagi."

Bagi manusia, kalimat tersebut dapat dipahami sebagai bentuk sindiran atau sentimen negatif. Namun, komputer tanpa pemahaman konteks akan melihat kata "luar biasa" sebagai kata positif dan berpotensi salah mengklasifikasikan kalimat tersebut sebagai opini positif. Fenomena ini menunjukkan bahwa tantangan analisis teks tidak hanya berkaitan dengan mengenali kata, tetapi juga memahami makna, hubungan antar kata, serta tujuan komunikasi pengguna.

1. Data Teks Bersifat Tidak Terstruktur (*Unstructured Data*)

Salah satu alasan utama mengapa data teks sulit dianalisis adalah karena sebagian besar teks tidak memiliki struktur yang konsisten. Data terstruktur seperti tabel database memiliki format yang jelas, misalnya kolom nama, usia, tanggal, atau jumlah transaksi. Komputer dapat langsung membaca dan melakukan operasi terhadap data tersebut. Sebaliknya, data teks memiliki bentuk yang sangat beragam. Sebuah dokumen dapat terdiri dari:

- kalimat dengan panjang berbeda,
- variasi bahasa,
- simbol,
- singkatan,
- kesalahan penulisan,

- emoji,
- kombinasi angka dan huruf.

Sebagai contoh, pengguna media sosial dapat menulis "Produknya bagus bgt!!! worth it parah " atau "produk bagus banget, sangat direkomendasikan"

Kedua kalimat tersebut memiliki makna yang hampir sama, tetapi bentuk penulisannya berbeda. Bagi manusia, kedua teks tersebut mudah dipahami sebagai rekomendasi positif. Namun, bagi komputer kedua teks tersebut merupakan rangkaian karakter yang berbeda sehingga membutuhkan proses tambahan agar dapat memahami kesamaan makna.

Dalam bidang Text Mining, permasalahan ini diatasi melalui tahap text preprocessing, yaitu proses mengubah data teks mentah menjadi bentuk yang lebih terstruktur. Tahapan seperti tokenisasi, normalisasi teks, stemming, dan penghapusan kata tidak penting dilakukan agar teks dapat diproses oleh algoritma computer

2. Kompleksitas Bahasa Manusia

Bahasa manusia memiliki tingkat kompleksitas yang tinggi karena tidak hanya menyampaikan informasi secara literal, tetapi juga mengandung makna tersembunyi, emosi, dan konteks sosial. Komputer mengalami kesulitan karena sebuah kata atau kalimat dapat memiliki lebih dari satu makna. Kondisi ini disebut sebagai ambiguity atau ambiguitas bahasa.

Contoh:

"Dia pergi ke bank." Kata *"bank"* dapat memiliki dua arti:

- lembaga keuangan,
- tempat di tepi sungai.

Manusia dapat menentukan arti berdasarkan konteks kalimat. Jika kalimatnya *"Dia pergi ke bank untuk mengambil uang."* Manusia langsung memahami bahwa bank berarti lembaga keuangan.

Namun, komputer membutuhkan informasi tambahan untuk menentukan makna yang benar. Dalam NLP, permasalahan ini diselesaikan melalui pendekatan semantic analysis dan model bahasa berbasis konteks seperti Transformer. Model modern seperti BERT dan GPT mampu memahami hubungan antar kata berdasarkan posisi dan konteks penggunaannya dalam kalimat.

3. Bahasa Mengandung Emosi dan Opini Subjektif

Berbeda dengan data numerik yang bersifat objektif, sebagian besar data teks manusia mengandung opini, perasaan, dan pengalaman pribadi.

Contoh:

"Laptop ini cepat, tetapi baterainya mengecewakan." Kalimat tersebut memiliki dua informasi aspek performa → positif, dan aspek baterai → negatif.

Jika sistem hanya melihat kata "cepat", maka sistem dapat memberikan klasifikasi positif yang tidak tepat. Permasalahan ini menjadi tantangan utama dalam sentiment analysis atau analisis sentimen. Analisis sentimen modern tidak hanya melihat kata positif dan negatif, tetapi juga mempertimbangkan konteks kalimat, intensitas emosi, dan hubungan antar aspek.

Contoh:

"Film ini tidak buruk." Secara sederhana terdapat kata "buruk" yang bernilai negatif. Namun, keseluruhan kalimat sebenarnya menunjukkan opini netral atau sedikit positif.

4. Data Teks Memiliki Noise yang Tinggi

Data teks yang berasal dari dunia nyata sering mengandung banyak gangguan (*noise*) yang dapat menurunkan kualitas analisis. Contoh data media sosial "*gk nyangkaaa produknya bagus bgt wkwkw 🤪*" Permasalahannya pada kalimat tersebut yaitu, "*gk*" merupakan singkatan dari "*tidak*", huruf berulang, penggunaan bahasa informal, emoji, dan kesalahan penulisan. Data seperti ini tidak dapat langsung digunakan dalam algoritma machine learning karena dapat menyebabkan model salah memahami pola. Oleh karena itu, proses preprocessing menjadi tahap yang sangat penting. Beberapa metode yang digunakan:

- *spelling correction*,
- *normalization*,
- *slang conversion*,
- *emoji interpretation*,
- *noise filtering*.

C. TEXT MINING DAN NLP DALAM SATU EKOSISTEM

Perkembangan teknologi digital menyebabkan pertumbuhan data teks yang sangat besar dan terus meningkat setiap waktu. Informasi dalam bentuk teks muncul dari berbagai aktivitas manusia, seperti percakapan media sosial, dokumen akademik, laporan bisnis, berita online, ulasan pelanggan, email, hingga rekam medis digital. Namun, sebagian besar data tersebut bersifat tidak terstruktur (*unstructured data*) sehingga tidak dapat langsung dianalisis menggunakan metode pengolahan data tradisional.

Untuk mengubah data teks menjadi informasi yang bernilai, diperlukan sebuah ekosistem teknologi yang menggabungkan Natural Language Processing (NLP) dan Text Mining. Kedua bidang ini

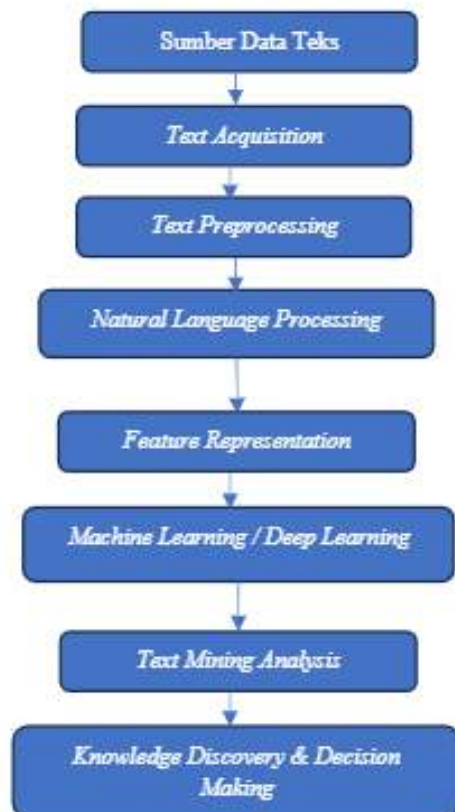
memiliki peran yang berbeda tetapi saling melengkapi. NLP berfungsi sebagai fondasi yang memungkinkan komputer memahami bahasa manusia, sedangkan Text Mining menggunakan kemampuan tersebut untuk menemukan pola, hubungan, dan pengetahuan baru dari kumpulan teks dalam jumlah besar.

Dalam sebuah ekosistem analisis teks modern, NLP bertindak sebagai lapisan pemahaman bahasa, sementara Text Mining bertindak sebagai proses eksplorasi dan ekstraksi informasi. Integrasi keduanya memungkinkan sistem melakukan berbagai tugas kompleks seperti klasifikasi dokumen, analisis sentimen, pencarian informasi, sistem rekomendasi, chatbot cerdas, hingga analisis prediksi berbasis teks.

1. Konsep Ekosistem Text Mining dan NLP

Ekosistem Text Mining dan NLP dapat dipahami sebagai rangkaian proses yang saling terhubung mulai dari pengumpulan data teks hingga menghasilkan informasi atau keputusan.

Secara sederhana, ekosistem tersebut terdiri dari beberapa lapisan utama:



Gambar 1. Ekosistem Text Mining dan NLP

Setiap lapisan memiliki fungsi berbeda. Data mentah yang awalnya sulit dipahami komputer secara bertahap diubah menjadi representasi yang dapat dianalisis dan digunakan untuk mendukung pengambilan keputusan.

2. Lapisan Pertama: Sumber Data Teks (Text Data Source)

Tahap awal dalam ekosistem Text Mining dan NLP adalah memperoleh data teks dari berbagai sumber. Sumber data dapat berasal dari:

- a. media sosial seperti komentar dan unggahan pengguna,
- b. website dan berita digital,
- c. dokumen perusahaan,

- d. email,
- e. forum diskusi,
- f. laporan penelitian,
- g. dokumen hukum,
- h. rekam medis,
- i. transkrip percakapan.

Sebagai contoh, perusahaan *e-commerce* dapat mengumpulkan jutaan ulasan pelanggan untuk memahami tingkat kepuasan pengguna terhadap produk tertentu.

Namun, data yang diperoleh biasanya masih berupa data mentah (*raw text*) yang memiliki banyak permasalahan seperti:

- a. format tidak seragam,
- b. kesalahan penulisan,
- c. penggunaan bahasa informal,
- d. simbol,
- e. emoji,
- f. campuran berbagai bahasa.

Oleh karena itu, data tersebut membutuhkan proses pengolahan sebelum dapat dianalisis lebih lanjut.

3. Lapisan Kedua: *Text Preprocessing* sebagai Penghubung Awal

Text preprocessing merupakan tahap penting yang menghubungkan data mentah dengan sistem NLP. Pada tahap ini, teks dibersihkan dan disiapkan agar dapat dipahami oleh algoritma komputer. Beberapa proses utama meliputi:

a. *Cleaning*

Cleaning bertujuan menghilangkan elemen yang tidak diperlukan seperti:

- tanda baca,
- karakter khusus,
- URL,

- simbol berlebihan.

Contoh:

Sebelum "Produk ini bagus banget!!! 😄🔥" Sesudah "produk ini bagus banget"

b. Case Folding

Case folding mengubah seluruh huruf menjadi bentuk yang sama.

Contoh "*Artificial Intelligence*" menjadi "*artificial intelligence*"

Hal ini mencegah komputer menganggap kata yang sama sebagai dua informasi berbeda.

4. Analisis Menggunakan Text Mining

Setelah data diproses menggunakan NLP, Text Mining melakukan analisis untuk menemukan informasi penting.

Beberapa penerapan utama yaitu:

a. Sentiment Analysis

Analisis sentimen digunakan untuk mengetahui opini atau emosi yang terdapat dalam teks.

Contoh:

Komentar "Produk ini sangat bagus dan pengirimannya cepat" dikategorikan sebagai Sentimen positif.

Penerapan:

- Analisis kepuasan pelanggan.
- Monitoring opini publik.
- Evaluasi layanan.

b. Text Classification

Text Classification bertujuan mengelompokkan dokumen berdasarkan kategori tertentu.

Contoh:

Email dapat diklasifikasikan *menjadi Spam, Promosi, Pesan penting*. Dalam bidang Pendidikan Mengelompokkan pertanyaan siswa, Mengidentifikasi tingkat kesulitan soal.

c. Named Entity Recognition (NER)

NER bertujuan mengenali informasi penting dalam teks seperti:

- Nama orang.
- Lokasi.
- Organisasi.
- Produk.
- Waktu.

Contoh:

"Apple meluncurkan iPhone terbaru di Amerika"

Sistem mengenali:

Apple → Organisasi.

iPhone → Produk.

Amerika → Lokasi.

D. PIPELINE PROYEK TEKS

Pipeline proyek teks merupakan rangkaian tahapan sistematis yang digunakan untuk mengelola, memproses, menganalisis, dan menghasilkan informasi dari data berbentuk teks. Dalam bidang Text Mining dan Natural Language Processing (NLP), pipeline ini berfungsi sebagai alur kerja (*workflow*) yang mengubah data teks mentah (*raw text*) yang tidak terstruktur menjadi informasi yang memiliki nilai analitis dan dapat digunakan untuk mendukung pengambilan keputusan.

Berbeda dengan data numerik yang umumnya sudah memiliki struktur tertentu, data teks memiliki karakteristik yang kompleks karena mengandung variasi bahasa, konteks, gaya penulisan, kesalahan ejaan, serta makna yang dapat berubah tergantung penggunaannya. Oleh karena itu, proyek berbasis teks membutuhkan pipeline yang terencana agar setiap tahap pemrosesan dapat menghasilkan data yang berkualitas.

Secara umum, pipeline proyek teks terdiri atas beberapa tahapan utama, yaitu pengumpulan data (*data acquisition*), pemahaman data (*data understanding*), preprocessing teks, ekstraksi fitur, pemodelan, evaluasi, dan implementasi hasil analisis.

1. Pengumpulan Data Teks (*Text Data Acquisition*)

Tahap pertama dalam pipeline proyek teks adalah mengumpulkan data yang akan dianalisis. Data teks dapat berasal dari berbagai sumber, baik berupa dokumen internal maupun sumber digital eksternal.

Sumber data teks dapat berupa:

- Media sosial seperti komentar, unggahan, dan percakapan pengguna.
- Ulasan produk pada marketplace.
- Artikel berita online.
- Dokumen akademik.
- Email perusahaan.
- Laporan pelanggan.
- Transkrip percakapan.
- Forum diskusi daring.

Pada tahap ini, kualitas data menjadi faktor penting karena hasil analisis sangat dipengaruhi oleh karakteristik data yang digunakan. Data yang terlalu sedikit, tidak relevan, atau memiliki banyak gangguan (*noise*) dapat menyebabkan model

menghasilkan prediksi yang kurang akurat. Contohnya, proyek analisis sentimen terhadap produk digital membutuhkan data berupa komentar pelanggan. Data tersebut harus mencerminkan opini pengguna secara nyata agar model mampu memahami pola sentimen positif maupun negatif.

2. Pemahaman dan Eksplorasi Data (*Text Data Understanding*)

Setelah data terkumpul, tahap berikutnya adalah memahami karakteristik data yang diperoleh. Tahapan ini bertujuan mengetahui struktur, kualitas, serta pola awal yang terdapat dalam kumpulan teks. Aktivitas yang dilakukan meliputi:

- Melihat jumlah dokumen atau teks yang tersedia.
- Mengidentifikasi bahasa yang digunakan.
- Mengetahui panjang rata-rata teks.
- Mengamati kata yang sering muncul.
- Mengidentifikasi data kosong atau duplikasi.
- Menganalisis distribusi kategori data.

Pada tahap ini, peneliti atau pengembang dapat menentukan metode analisis yang paling sesuai. Sebagai contoh, data komentar media sosial biasanya memiliki karakteristik:

- Banyak menggunakan singkatan.
- Mengandung emoji.
- Menggunakan bahasa informal.
- Memiliki kesalahan penulisan.

E. PENDEKATAN RULE-BASED, STATISTIK, DAN NEURAL

Dalam perkembangan Natural Language Processing (NLP) dan Text Mining, terdapat tiga pendekatan utama yang digunakan untuk membuat komputer mampu memahami, mengolah, dan menganalisis bahasa manusia, yaitu pendekatan berbasis aturan (*Rule-based*

Approach), pendekatan statistik (*Statistical Approach*), dan pendekatan berbasis jaringan saraf (*Neural Approach*).

Ketiga pendekatan tersebut menunjukkan perkembangan paradigma dalam pengolahan bahasa alami. Pendekatan awal lebih mengandalkan aturan linguistik yang dibuat secara manual oleh manusia, kemudian berkembang menjadi pendekatan statistik yang memanfaatkan pola dari data, hingga pendekatan neural yang menggunakan jaringan saraf tiruan untuk memahami konteks bahasa secara lebih mendalam.

Perubahan dari pendekatan Rule-based menuju Neural menunjukkan pergeseran dari sistem yang bergantung pada pengetahuan manusia (*knowledge-driven*) menuju sistem yang belajar secara otomatis dari data (*data-driven*).

1. Pendekatan Rule-based (*Rule-based Approach*)

Pendekatan Rule-based merupakan metode awal dalam NLP yang menggunakan aturan linguistik yang dibuat secara eksplisit oleh manusia. Dalam pendekatan ini, komputer tidak belajar dari data, tetapi mengikuti sekumpulan aturan yang telah dirancang oleh ahli bahasa atau pengembang sistem. Aturan tersebut biasanya berbentuk pola tertentu seperti:

- aturan tata bahasa (*grammar rules*);
- daftar kata (*lexicon*);
- aturan sintaksis;
- aturan pencocokan pola (*pattern matching*).

Sistem akan mengambil keputusan berdasarkan apakah suatu teks sesuai dengan aturan yang telah ditentukan.

Contoh sederhana:

Aturan analisis sentimen:

- Jika teks mengandung kata "bagus", "baik", atau "luar biasa", maka dikategorikan positif.

- Jika teks mengandung kata "buruk", "jelek", atau "kecewa", maka dikategorikan negatif.

Kalimat: "Produk ini sangat bagus" akan dikenali sebagai sentimen positif karena terdapat kata "bagus".

Karakteristik Pendekatan Rule-based

- a. Menggunakan aturan eksplisit
Sistem bekerja berdasarkan logika yang dibuat manusia.
Contoh:
IF kata = "murah"
THEN sentimen = positif
- b. Tidak membutuhkan data pelatihan besar
Berbeda dengan metode Machine Learning, pendekatan ini tidak membutuhkan ribuan atau jutaan contoh data untuk belajar. Pengembang cukup membuat aturan yang sesuai dengan kebutuhan sistem.
- c. Mudah dijelaskan (*interpretable*)
Salah satu keunggulan utama Rule-based adalah keputusan sistem dapat dijelaskan. Jika sistem memberikan hasil tertentu, pengembang dapat melihat aturan mana yang menyebabkan keputusan tersebut.

2. Pendekatan Statistik (*Statistical Approach*)

Pendekatan statistik muncul sebagai perkembangan dari keterbatasan metode Rule-based. Pendekatan ini menggunakan teori probabilitas dan data untuk menemukan pola bahasa. Berbeda dengan Rule-based yang menggunakan aturan manual, pendekatan statistik memungkinkan komputer mempelajari hubungan antar kata berdasarkan data teks yang tersedia. Prinsip utama pendekatan statistik Adalah Bahasa dapat dipahami

melalui pola kemunculan kata dan hubungan probabilitas antar elemen bahasa.

Contoh:

Jika dalam banyak dokumen kata "kopi" sering muncul Bersama "minuman", "cangkir", "rasa", maka sistem dapat mempelajari bahwa kata tersebut memiliki hubungan dengan konteks minuman.

Metode dalam Pendekatan Statistik

a. N-gram Model

N-gram merupakan metode yang memprediksi kata berdasarkan kata sebelumnya.

Contoh:

Kalimat "Saya minum..." Kemungkinan kata berikutnya (kopi, air, atau teh)

Sistem menghitung kemungkinan berdasarkan data sebelumnya.

Jenis N-gram:

- Unigram: satu kata.
- Bigram: dua kata.
- Trigram: tiga kata.

b. Hidden Markov Model (HMM)

HMM merupakan model probabilistik yang digunakan untuk memahami urutan kata.

Penerapan:

- Part-of-Speech Tagging.
- Speech Recognition.
- Named Entity Recognition.

Contoh:

Kalimat "Bank Indonesia mengeluarkan kebijakan baru" Sistem menentukan bahwa "Bank Indonesia" merupakan nama institusi, bukan tempat menyimpan uang.

c. Naive Bayes

Naive Bayes merupakan algoritma statistik yang sering digunakan untuk klasifikasi teks.

Penerapan:

- Deteksi spam email.
- Klasifikasi berita.
- Analisis sentimen.

Cara kerja

Model menghitung kemungkinan suatu teks termasuk dalam kategori tertentu berdasarkan pola kata.

3. Pendekatan Neural (*Neural Approach*)

Pendekatan Neural merupakan perkembangan terbaru dalam NLP yang menggunakan konsep Artificial Neural Network atau jaringan saraf tiruan. Pendekatan ini memungkinkan komputer belajar representasi bahasa secara otomatis melalui proses pelatihan menggunakan data dalam jumlah besar. Berbeda dengan pendekatan statistik yang menggunakan fitur buatan manusia, Neural Approach mampu mempelajari fitur secara otomatis (*automatic feature learning*).

F. CONTOH MASALAH

Contoh penggunaan mencakup klasifikasi tiket layanan, moderasi konten, deteksi spam, analisis opini, pencarian dokumen, ekstraksi entitas, ringkasan dokumen, pencocokan CV, analisis literatur, penerjemahan, chatbot, serta pemantauan isu publik. Setiap aplikasi

memiliki konsekuensi berbeda terhadap biaya kesalahan. False negative pada moderasi risiko tinggi dapat lebih berbahaya daripada false positive, sedangkan pada pencarian dokumen trade-off precision dan recall dapat diatur sesuai kebutuhan pengguna.

Pemilihan dataset dan skema label harus merepresentasikan situasi penggunaan. Dataset ulasan film tidak otomatis cocok untuk sentimen layanan kesehatan. Model yang diuji pada bahasa formal dapat gagal pada media sosial yang mengandung singkatan, emoji, code-switching, dan ejaan tidak baku. Generalisasi lintas domain harus diuji, bukan diasumsikan.

Pendekatan	Kekuatan	Keterbatasan	Contoh
<i>Rule-based</i>	Transparan, mudah dikontrol	Sulit menangani variasi besar	Regex, kamus, grammar
<i>Machine learning klasik</i>	Cepat, baseline kuat, interpretable	Butuh fitur/representasi	NB, logistic regression, SVM
<i>Deep learning</i>	<i>Representation learning</i>	Butuh data/komputasi	CNN, LSTM
<i>Transformer</i>	Konteks kaya, transfer learning	Biaya lebih tinggi	BERT, T5
<i>LLM + retrieval</i>	Generasi fleksibel, akses dokumen	Hallucination dan governance	RAG, agentic workflow terbatas

G. LATIHAN DAN DISKUSI

1. Berikan tiga contoh masalah text mining dan jelaskan *output* keputusan yang dihasilkan.
2. Mengapa model paling kompleks tidak selalu menjadi pilihan terbaik?
3. Bandingkan *rule-based*, *machine learning*, dan Transformer pada kasus klasifikasi tiket.

BAB 2

KORPUS, AKUISISI DATA, ANOTASI, DAN TATA KELOLA

A. CAPAIAN PEMBELAJARAN

Mahasiswa mampu mempelajari dan memahami

1. Merancang korpus yang sesuai dengan populasi sasaran.
2. Menyusun pedoman anotasi dan mengukur konsistensi label.
3. Menilai aspek lisensi, privasi, bias, dan provenance data.

B. KORPUS DAN REPRESENTATIVITAS

Korpus (*corpus*) merupakan salah satu komponen fundamental yang menentukan kualitas suatu sistem analisis teks. Korpus menyediakan kumpulan data bahasa yang menjadi sumber pembelajaran, pengujian, maupun evaluasi bagi berbagai metode NLP.

Kualitas suatu proyek NLP tidak hanya ditentukan oleh algoritma atau model yang digunakan, tetapi juga sangat bergantung pada kualitas data yang menjadi dasar analisis. Data teks yang digunakan harus mampu menggambarkan karakteristik bahasa atau fenomena yang ingin dikaji. Konsep inilah yang berkaitan dengan representativitas korpus, yaitu sejauh mana sebuah kumpulan teks mampu mewakili populasi bahasa atau fenomena yang menjadi tujuan penelitian. Dengan kata lain, korpus yang baik bukan sekadar

memiliki jumlah data yang besar, tetapi harus memiliki struktur, variasi, dan distribusi yang sesuai dengan konteks penggunaannya.

1. Pengertian Korpus (*Corpus*)

Korpus adalah kumpulan teks yang dikumpulkan secara sistematis dan disimpan dalam format digital untuk tujuan analisis bahasa. Data dalam korpus dapat berupa teks tertulis maupun transkripsi bahasa lisan yang telah melalui proses pengolahan tertentu sehingga dapat dianalisis menggunakan metode komputasi. Dalam NLP, korpus berfungsi sebagai sumber data untuk memahami pola penggunaan bahasa manusia.

Contoh korpus:

- Kumpulan artikel berita untuk analisis topik.
- Kumpulan komentar pengguna media sosial untuk analisis sentimen.
- Kumpulan dokumen hukum untuk sistem pencarian informasi hukum.
- Kumpulan percakapan pelanggan untuk pengembangan chatbot.
- Kumpulan karya ilmiah untuk analisis literatur.

2. Fungsi Korpus dalam Text Mining dan NLP

Korpus memiliki beberapa fungsi utama dalam pengembangan sistem berbasis bahasa.

a. sebagai Data Pelatihan (*Training Data*)

Korpus digunakan untuk melatih model agar mampu mengenali pola bahasa.

Contoh:

Model analisis sentimen diberikan ribuan komentar yang sudah memiliki label:

- Positif.
- Negatif.

DAFTAR PUSTAKA

- Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python*. O'Reilly Media.
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of NAACL-HLT*.
- Hu, E. J., et al. (2022). LoRA: Low-Rank Adaptation of Large Language Models. *International Conference on Learning Representations*.
- Jurafsky, D., & Martin, J. H. (2026). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models* (3rd ed., online manuscript released January 6, 2026). Stanford University.
- Koto, F., Rahimi, A., Lau, J. H., & Baldwin, T. (2020). IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP. *Proceedings of COLING*.
- Lewis, P., et al. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *Advances in Neural Information Processing Systems*.
- Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.
- Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *ICLR Workshop*.
- Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global Vectors for Word Representation. *Proceedings of EMNLP*.

- Raffel, C., et al. (2020). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21(140).
- Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *Proceedings of EMNLP-IJCNLP*.
- Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5), 513–523.
- Vaswani, A., et al. (2017). Attention Is All You Need. *Advances in Neural Information Processing Systems*.
- Wilie, B., et al. (2020). IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding. *Proceedings of ACL-IJCNLP*.
- Winata, G. I., et al. (2023). NusaX: Multilingual Parallel Sentiment Dataset for 10 Indonesian Local Languages. *Proceedings of EACL*.
- NLTK Project. *Natural Language Toolkit Documentation and NLTK Book*. Accessed 2026.
- scikit-learn Developers. *Feature Extraction and Working with Text Documents documentation*. Accessed 2026.
- explosion.ai. *spaCy Usage Documentation: Linguistic Features and Processing Pipelines*. Accessed 2026.
- Hugging Face. *Transformers Documentation: BERT, T5, fine-tuning, text classification, and summarization*. Accessed 2026.
- PyTorch. *Transformer and NLP Tutorials documentation*. Accessed 2026.

GLOSARIUM

Istilah	Definisi
<i>Annotation</i>	Proses pemberian label atau struktur pada data teks.
<i>Attention</i>	Mekanisme pembobotan hubungan antarposisi dalam sequence.
<i>BERT</i>	Model encoder Transformer pralatih dengan objective masked language modeling.
<i>BM25</i>	Fungsi ranking lexical retrieval berbasis term frequency dan document length.
<i>Corpus</i>	Koleksi teks yang digunakan untuk analisis atau pemodelan.
<i>Data leakage</i>	Masuknya informasi yang seharusnya tidak tersedia pada tahap training
<i>Dense retrieval</i>	Pencarian berdasarkan kedekatan embedding query dan dokumen.
<i>Embedding</i>	Representasi numerik dense untuk token, kalimat, atau dokumen.
<i>Fine-tuning</i>	Adaptasi model pralatih dengan melanjutkan training pada task target.
<i>Hallucination</i>	Keluaran generatif yang tidak didukung fakta atau sumber yang tersedia.
<i>Information retrieval</i>	Proses mencari dan memeringkat dokumen yang relevan terhadap query.
<i>Lemmatization</i>	Normalisasi kata ke bentuk lemma berdasarkan analisis morfologi.
<i>LLM</i>	Large Language Model, model bahasa berskala besar yang umumnya bersifat generatif.

Istilah	Definisi
<i>NER</i>	Named Entity Recognition, identifikasi span entitas dan kategorinya.
<i>N-gram</i>	Urutan n unit berdekatan, misalnya unigram, bigram, atau trigram.
<i>Prompt injection</i>	Upaya memengaruhi instruksi model melalui input atau konten yang tidak terpercaya.
<i>RAG</i>	Retrieval-Augmented Generation, arsitektur yang menggabungkan retrieval dan generation.
<i>Stemming</i>	Reduksi bentuk kata ke stem menggunakan aturan atau algoritma.
<i>TF-IDF</i>	Pembobotan term berdasarkan frekuensi lokal dan inverse document frequency.
<i>Tokenization</i>	Pemecahan teks menjadi unit token.
<i>Topic modeling</i>	Pemodelan struktur topik laten dalam koleksi dokumen.
<i>Transformer</i>	Arsitektur neural berbasis attention yang menjadi fondasi banyak model bahasa modern
<i>Vector search</i>	Pencarian nearest neighbor pada ruang embedding.

PROFIL PENULIS



Nama Alfiansyah Hasibuan. Menyelesaikan pendidikan Tingkat SD, SMP, SMA Negeri di kecamatan darul makmur. Untuk pendidikan tinggi, penulis menyelesaikan Program Diploma 3 pada tahun 2011 dan Sarjana pada tahun 2012 di Sekolah Tinggi Manajemen Informatika dan Komputer LOGIKA di kota medan sumatera Utara, dan menyelesaikan program magister pada tahun 2019 di Program Studi Teknik Informatika Universitas Sumatera Utara. Saat ini penulis sebagai tenaga pengajar di Program Studi Teknik Informatika Fakultas Teknik di Universitas Negeri Manado.



Abraham Yano Suharmanto, S.Kom., M.Kom adalah dosen di Program Studi Informatika Kesehatan, Sekolah Tinggi Ilmu Kesehatan Gunung Maria Tomohon. Menyelesaikan pendidikan S-1 di Teknik Informatika, Universitas Sam Ratulangi Manado dan S-2 Magister Sistem Informasi (MSI) di Universitas Diponegoro Semarang dengan fokus kepakaran pada Teknik Komunikasi Jaringan dan Sistem Informasi.



Prof. Dr. Ing. Drs. Parabelem Tinno Dolf Rompas, MT. Adalah dosen di Fakultas Teknik Universitas Negeri Manado (UNIMA). Menyelesaikan pendidikan S-1 di IKIP Negeri Manado, S-2 Di Universitas Gadjah Mada, dan S-3 di Universite Aix-Marseille III, Prancis dengan fokus pada Mekanika Fluida dan Modelisasi Numerik.



Nama Sapriadi, Lahir di Medan. Menyelesaikan pendidikan Tingkat SD, SMP, SMA Negeri di Kota Medan. Untuk pendidikan tinggi, penulis menyelesaikan Program Sarjana pada tahun 2014 di Sekolah Tinggi Manajemen Informatika dan Komputer LOGIKA di kota medan sumatera Utara, dan menyelesaikan program magister pada tahun 2019 di Program Studi Teknik Informatika Universitas Sumatera Utara. Saat ini penulis sebagai tenaga pengajar di Institut Kesehatan Helvetia Medan.

TEXT MINING dan NATURAL LANGUAGE PROCESSING

Text Mining merupakan proses menemukan pola, informasi, dan pengetahuan dari kumpulan data teks melalui serangkaian tahapan sistematis. Dalam ekosistem NLP, proses tersebut dimulai dari akuisisi data, pembentukan korpus, pemeriksaan kualitas data melalui audit sebelum cleaning, hingga proses preprocessing seperti tokenization, case folding, stopword removal, stemming, lemmatization, normalisasi bahasa informal, serta penanganan negasi, emoji, URL, dan entitas khusus. Tahapan ini bertujuan menghasilkan data teks yang lebih bersih dan konsisten sehingga dapat digunakan untuk proses analisis berikutnya.

Kualitas data menjadi faktor penting dalam keberhasilan proyek NLP. Oleh karena itu, pemahaman mengenai korpus dan representativitas, distribusi panjang teks, serta deteksi outlier diperlukan agar data yang digunakan mampu menggambarkan fenomena sebenarnya. Selain itu, proses anotasi dan penyusunan skema label menjadi bagian penting terutama pada pendekatan pembelajaran terawasi (*supervised learning*) karena kualitas label akan memengaruhi kemampuan model dalam mengenali pola bahasa.

Dalam proses representasi teks, terdapat berbagai pendekatan yang berkembang dari metode statistik hingga pendekatan berbasis kecerdasan buatan. Metode awal seperti Bag of Words, Document-Term Matrix, N-gram, dan TF-IDF mengubah teks menjadi bentuk numerik berdasarkan frekuensi dan distribusi kata. Pendekatan tersebut efektif untuk berbagai tugas klasifikasi, tetapi memiliki keterbatasan karena belum mampu memahami hubungan makna, konteks, dan struktur bahasa secara mendalam.

Salah satu penerapan penting dalam NLP adalah analisis sentimen, yaitu proses mengidentifikasi opini, emosi, atau sikap seseorang terhadap suatu objek tertentu. Analisis sentimen tidak hanya melihat apakah suatu teks bernilai positif, negatif, atau netral, tetapi juga dapat dikembangkan menjadi analisis berbasis target (*target sentiment* atau Aspect-Based Sentiment Analysis) untuk mengetahui sentimen terhadap aspek tertentu dari sebuah objek. Contohnya, sebuah ulasan dapat memberikan sentimen positif terhadap kualitas kamera, tetapi negatif terhadap daya tahan baterai.

Perkembangan terbaru dalam NLP membawa perubahan besar melalui Deep Learning dan Representation Learning. Berbeda dengan pendekatan fitur manual yang membutuhkan manusia untuk menentukan karakteristik data, representation learning memungkinkan model mempelajari fitur secara otomatis dari teks. Teknologi seperti Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM), Convolutional Neural Network (CNN), hingga Transformer seperti BERT dan GPT mampu memahami konteks bahasa secara lebih mendalam melalui mekanisme perhatian (*attention mechanism*).



CV. Tahta Media Group
Sukoharjo, Jawa Tengah
Web : www.tahtamedia.co.id
Ig : [tahtamedia](https://www.instagram.com/tahtamedia)
Telp/WA : +62 896 5427 3996

