KLASIFIKASI KETERAMPILAN KERJA MENGGUNAKAN METODE TF-IDF DAN DECISION TREE PADA DATA LOWONGAN KERJA LINKEDIN

THESIS

Oleh:

MOHAMUD AHMED

NIM. 210605210007

PROGRAM STUDI MAGISTER INFORMATIKA
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI MAULANA MALIK IBRAHIM
MALANG
2025

KLASIFIKASI KETERAMPILAN KERJA MENGGUNAKAN METODE TF-IDF DAN DECISION TREE PADA DATA LOWONGAN KERJA LINKEDIN

THESIS

Diajukan kepada:
Universitas Islam Negeri Maulana Malik Ibrahim Malang
Untuk memenuhi Salah Satu Persyaratan dalam
Memperoleh Gelar Magister Komputer (M.Kom)

Oleh:

MOHAMUD AHMED

NIM. 210605210007

PROGRAM STUDI MAGISTER INFORMATIKA
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI MAULANA MALIK IBRAHIM MALANG
2025

KLASIFIKASI KETERAMPILAN KERJA MENGGUNAKAN METODE TF-IDF DAN DECISION TREE PADA DATA LOWONGAN KERJA LINKEDIN

THESIS

Diajukan kepada:
Fakultas Sains dan Teknologi
Universitas Islam Negeri Maulana Malik Ibrahim Malang
Untuk memenuhi Salah Satu Persyaratan dalam
Memperoleh Gelar Magister Komputer (M.Kom)

Oleh:

MOHAMUD AHMED

NIM. 210605210007

PROGRAM STUDI MAGISTER INFORMATIKA
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI MAULANA MALIK IBRAHIM MALANG
2025

KLASIFIKASI KETERAMPILAN KERJA MENGGUNAKAN METODE TF-IDF DAN DECISION TREE PADA DATA LOWONGAN KERJA LINKEDIN

THESIS

Oleh:
MOHAMUD AHMED
NIM. 210605210007

Telah diperiksa dan disetujui untuk di uji:
Tanggal 18 Juni 2025

Pembimbing I	Pembimbing II
Prof. Dr. Suhartono, M.Kom	Dr. M. Imamudin Lc, MA
NIP.19680519 200312 1 001	NIP 19740602 200901 1 010

Mengetahui,
Ketua Program Studi Magister Informatika
Fakultas Sains dan Teknologi
Universitas Islam Maulana Malik Ibrahim Malang

Dr. Fachrul Kurniawan
NIP. 19740424 200901 1 008

KLASIFIKASI KETERAMPILAN KERJA MENGGUNAKAN METODE TF-IDF DAN DECISION TREE PADA DATA LOWONGAN KERJA LINKEDIN

THESIS

Oleh:
MOHAMUD AHMED
NIM. 210605210007

Telah Dipertahankan di Depan Dewan Penguji Thesis
dan Dinyatakan Diterima Sebagai Salah Satu Persyaratan
Untuk Memperoleh Gelar Magister Komputer (M.Kom)
Tanggal 18 Juni 2025

Susunan Dewan Penguji			**Tanda Tangan**
Penguji I	:	Dr. Totok Chamidy, M.Kom NIP 19691222 200604 1 001	( )
Penguji II	:	Dr. Agung Teguh Wibowo Almais, M.T NIP 1986030 1202321 1 016	( )
Pembimbing I	:	Prof. Dr. Suhartono, M.Kom NIP.19680519 200312 1 001	( )
Pembimbing II	:	Dr. M. Imamudin Lc, MA NIP 19740602 200901 1 010	

Mengetahui,
Ketua Program Studi Magister Informatika
Fakultas Sains dan Teknologi
Universitas Islam Maulana Malik Ibrahim Malang

Dr. Cahyo Crysdian
NIP. 19740424 200901 1 008

PERNYATAAN KEASLIAN TULISAN

Saya yang bertanda tangan dibawah ini:

Nama : Mohamud Ahmed Mohamed

NIM 210605210007

Program Studi : Magister Informatika

Fakultas : Sains dan Teknologi

Menyatakan dengan sebenarnya bahwa Thesis yang saya tulis ini benar-benar merupakan hasil karya saya sendiri, bukan merupakan pengambilalihan data, tulisan atau pikiran orang lain yang saya akui sebagai tulisan atau pikiran saya sendiri, kecuali dengan mencantumkan sumber cuplikan pada daftar pustaka.

Apabila dikemudian hari terbukti atau dapat dibuktikan Thesis ini hasil jiplakan, maka saya bersedia menerima sanksi atas perbuatan tersebut.

Malang, 18 Juni 2025

Yang Membuat Pernyataan

Moahamud Ahmed

NIM. 210605210007

MOTTO

“ Stay Hungry, Stay Foolish “

Steve Job

PERSEMBAHAN

Atas berkah dan rahmat Allah SWT, Thesis ini bisa saya selesaikan dengan lengkap dan baik sesuai dengan arahan bapak ibu pembimbing, oleh karena itu hasil kerja keras atas keberhasilan dari penulisan Thesis ini merupakan dukungan dari beberapa pihak yaitu :

1. Keluarga kecil tercinta antara lain istri
2. Keluarga besar, yaitu ibu, ayah, ayah mertua, kakek, nenek, paman, kakak sepupu dan saudara-saudara yang lain.
3. Bapak ibu dosen Magister Informatika Fakultas Sains dan Teknologi Universitas Islam Negeri Maulana Malik Ibrahim Malang.
4. Teman-teman angkatan Magister Informatika Fakultas Sains dan Teknologi Universitas Islam Negeri Maulana Malik Ibrahim Malang.
5. Teman-teman mahasiswa Magister Informatika Fakultas Sains dan Teknologi Universitas Islam Negeri Maulana Malik Ibrahim Malang.
6. Serta rekan-rekan yang tidak mungkin disebutkan satu persatu atas dukungan terselesaikannya Thesis ini.

KATA PENGANTAR

Assalamu'alaikum Warahmatullahi Wa barakatuh

Syukur *Alhamdulillah* penulis haturkan kehadirat Allah SWT yang telah melimpahkan Rahmat dan Hidayah-Nya, sehingga penulis dapat menyelesaikan studi di Program Studi Magister Informatika Fakultas Sains dan Teknologi Universitas Islam Negeri Maulana Malik Ibrahim Malang sekaligus menyelesaikan Thesis ini dengan baik.

Selanjutnya penulis haturkan ucapan terima kasih seiring do'a dan harapan jazakumullah ahsanal jaza' kepada semua pihak yang telah membantu terselesaikannya Thesis ini. Ucapan terima kasih ini penulis sampaikan kepada:

1. Prof. Dr. Suhartono, M.Kom., dan Dr. M. Imamudin Lc, MA selaku dosen pembimbing Thesis, yang telah banyak memberikan pengarahan dan pengalaman yang berharga.
2. Segenap civitas akademika Program Studi Magister Informatika, terutama seluruh Bapak / Ibu dosen, terima kasih atas segenap ilmu dan bimbingannya.
3. Keluarga tercinta yang senantiasa memberikan do'a dan semangat
4. Semua rekan-rekan seperjuangan yang ikut mendukung dan membantu.

Penulis menyadari bahwa dalam penyusunan Thesis ini masih terdapat kekurangan dan penulis berharap semoga Thesis ini bisa memberikan manfaat kepada para pembaca khususnya bagi penulis secara pribadi. Amiinn Yaa Rabbal Alamin.

Wasalamu'alaikum Warahmatullahi Wa barakatuh

Malang, 18 Juni 2025

Penulis

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

ABSTRAK

Mohamud Ahmed. 2025. Klasifikasi Keterampilan Kerja menggunakan Metode TF-IDF dan*Decision Tree* pada Data Lowongan Kerja LinkedIn. Thesis Program Studi Magister Informatika Fakultas Sains dan Teknologi Universitas Islam Negeri Maulana Malik Ibrahim Malang. Pembimbing: (I) Prof. Dr. Suhartono, M.Kom., (II) Dr. M. Imamudin Lc, MA.,

Kata Kunci: TF-IDF, Decision Tree, Keterampilan Kerja, Lowongan Pekerjaan, Text Mining, Industri 4.0

Penelitian ini bertujuan untuk menganalisis tren keterampilan kerja di sektor Teknologi Informasi (IT) dengan memanfaatkan data lowongan pekerjaan sintetis yang menyerupai format LinkedIn menggunakan pendekatan text mining. Metode TF-IDF diterapkan untuk mengekstraksi fitur kata kunci penting dari deskripsi pekerjaan yang bersifat tidak terstruktur, sementara algoritma Decision Tree digunakan untuk mengklasifikasikan jenis pekerjaan berdasarkan fitur yang diperoleh. Data yang digunakan meliputi 100 entri lowongan pekerjaan berbahasa campuran Indonesia dan Inggris, dengan proses preprocessing teks yang komprehensif untuk memastikan kualitas data. Hasil penelitian menunjukkan bahwa kombinasi TF-IDF dan Decision Tree efektif dalam mengidentifikasi keterampilan utama dan mengelompokkan jenis pekerjaan secara akurat dan mudah diinterpretasikan. Berdasarkan analisis, kategori pekerjaan Data Engineer menjadi yang paling banyak diminati, dengan kata kunci utama seperti “data”, “experi”, “work”, “team”, dan “product” yang menggambarkan kebutuhan keterampilan teknis dan kolaboratif. Model Decision Tree mencapai akurasi 80,3%, khususnya baik dalam mengklasifikasikan Data Analyst. Visualisasi seperti Word Cloud dan feature importance plot memberikan gambaran intuitif mengenai kebutuhan keterampilan yang dapat dimanfaatkan oleh pencari kerja, penyusun kurikulum, dan perusahaan rekrutmen. Kesimpulannya, penelitian ini membuktikan bahwa penggunaan metode TF-IDF dan Decision Tree mampu mengotomatisasi klasifikasi keterampilan kerja dari data lowongan pekerjaan secara efektif, sehingga mendukung pengambilan keputusan berbasis data dalam dunia ketenagakerjaan di era digital dan revolusi industri 4.0..

ABSTRACT

Mohamud Ahmed. 2025. Klasifikasi Keterampilan Kerja menggunakan Metode TF- ID dan*Decision Tree* pada Data Lowongan Kerja LinkedIn. Thesis Program Studi Magister Informatika Fakultas Sains dan Teknologi Universitas Islam Negeri Maulana Malik Ibrahim Malang. Pembimbing: (I) Prof. Dr. Suhartono, M.Kom., (II) Dr. M. Imamudin Lc, MA.,

.

Keywords: TF-IDF, Decision Tree, Job Skills, Job Vacancies, Text Mining, Industry 4.0

This study aims to analyze skill trends in the Information Technology (IT) sector by utilizing synthetic job vacancy data resembling LinkedIn format through a text mining approach. The TF-IDF method was applied to extract important keyword features from unstructured job descriptions, while the Decision Tree algorithm was used to classify job types based on the extracted features.The dataset consists of 100 job listings in mixed Indonesian and English languages, with comprehensive text preprocessing to ensure data quality. The results indicate that the combination of TF-IDF and Decision Tree is effective in identifying key skills and categorizing job types accurately and interpretably. Data Engineer emerged as the most sought-after job category, with dominant keywords such as “data,” “experi,” “work,” “team,” and “product” reflecting the need for both technical and collaborative skills. The Decision Tree model achieved an accuracy of 80.3%, performing particularly well in classifying Data Analyst positions. Visualizations, including Word Cloud and feature importance plots, provide intuitive insights into skill demands that can benefit job seekers, curriculum developers, and recruitment companies. In conclusion, this study demonstrates that employing TF-IDF and Decision Tree methods can effectively automate the classification of job skills from vacancy data, thereby supporting data-driven decision-making in the workforce amidst the digital era and Industry 4.0 revolution..

ملخص

محمود أحمد، 2025متصنيف المهارات الوظيفية باستخدام طريقة TF-IDF وخوارزمية Decision Tree على بيانات إعلانات الوظائف في .LinkedIn رسالة ماجستير، برنامج ماجستير علوم الحاسوب، كلية العلوم والتكنولوجيا، الجامعة الإسلامية الحكومية مولانا مالك إبراهيم مالانج .المشرفان: (1) الأستاذ الدكتور سوهارتونو، (2)الدكتور محمد إمام الدين،

الكلمات المفتاحية: TF-IDF ؛ شجرة القرار، مهارات العمل، الوظائف الشاغرة، تنقيب النصوص، الثورة الصناعية الرابعة.

تهدف هذه الدراسة إلى تحليل اتجاهات المهارات في قطاع تكنولوجيا المعلومات من خلال استخدام بيانات وظائف صناعية تحاكي تنسيق منصةLinkedIn ، وذلك باستخدام نهج تنقيب النصوص. تم تطبيق طريقة TF-IDF لاستخلاص الكلمات المفتاحية المهمة من أوصاف الوظائف غير المهيكلة، في حين استُخدم خوارزم Decision Treeلتصنيف أنواع الوظائف استنادًا إلى الميزات المستخرجة.

يتكون مجموعة البيانات من 100 إعلان وظيفة بلغتين: الإندونيسية والإنجليزية، مع إجراء معالجة نصوص شاملة لضمان جودة البيانات. تشير النتائج إلى أن الجمع بين TF-IDF وخوارزم Decision Tree فعال في تحديد المهارات الأساسية وتصنيف أنواع الوظائف بدقة ووضوح. وبرزت وظيفة مهندس بيانات (Data Engineer)كأكثر الوظائف طلبًا، مع كلمات مفتاحية بارزة مثل"data" ، "experi"، "work"، "team"، و"product"، مما يعكس الحاجة إلى المهارات التقنية والقدرة على العمل الجماعي.

حقق نموذج Decision Tree دقة بلغت 80.3%، وكان أداؤه متميزًا خصوصًا في تصنيف وظائف محللي البيانات .(Data Analyst) توفر التصورات البيانية مثل سحابة الكلمات ومخططات أهمية الميزات رؤى بديهية حول متطلبات المهارات، مما يعود بالفائدة على الباحثين عن عمل، ومطوري المناهج الدراسية، وشركات التوظيف. وفي الختام، تُظهر هذه الدراسة أن استخدام طريقتي TF-IDF و Decision Treeيمكن أن يساهم بفعالية في أتمتة تصنيف المهارات المطلوبة في سوق العمل من خلال بيانات الوظائف، مما يدعم اتخاذ قرارات قائمة على البيانات في ظل التحول الرقمي وثورة الصناعة 4.0.

BAB I

PENDAHULUAN

1.1 Latar Belakang

Indonesia merupakan salah satu negara yang memiliki pertumbuhan pengguna media sosial terbesar di dunia dalam satu tahun, yaitu 20 juta pengguna pada tahun 2019 (Kemp, 2019). Pertumbuhan Media Sosial terjadi dibarengin dengan meningkatnya jumlah pengguna internet di Indonesia yang ditunjukan dengan era keterbukaan informasi yang semakin tersebar luas, sehingga memudahkan masyarakat untuk memperoleh informasi (Pratama & Tjahyanto, 2021). Perkembangan teknologi informasi telah membawa perubahan besar dalam dunia kerja, terutama dalam proses rekrutmen tenaga kerja. Salah satu platform yang saat ini banyak digunakan oleh perusahaan dan pencari kerja adalah LinkedIn. LinkedIn menyediakan berbagai data lowongan kerja yang bersifat tidak terstruktur namun memiliki nilai informasi yang tinggi.

Di era revolusi industri 4.0 dan transformasi digital, dunia kerja mengalami perubahan yang sangat cepat dan dinamis. Kemajuan teknologi informasi dan komunikasi telah mendorong munculnya berbagai jenis pekerjaan baru serta keterampilan yang dibutuhkan juga berubah seiring waktu. Oleh karena itu, memahami keterampilan kerja yang paling relevan dan dibutuhkan oleh pasar menjadi hal yang krusial bagi para pencari kerja, perusahaan, dan lembaga pendidikan maupun pelatihan. Informasi tentang keterampilan kerja yang akurat dan terstruktur dapat membantu para pencari kerja menyesuaikan diri dengan

kebutuhan industri serta membantu perusahaan dalam proses rekrutmen tenaga kerja yang tepat (Nissa et al., 2025).

LinkedIn sebagai salah satu platform profesional terbesar di dunia, menyediakan data lowongan kerja yang sangat kaya dan beragam. Lowongan kerja tersebut biasanya memuat informasi penting terkait posisi yang dibuka, persyaratan keterampilan, pengalaman, dan kualifikasi yang dibutuhkan. Namun, data lowongan kerja yang tersebar dan tidak terstruktur dalam bentuk teks bebas menjadi tantangan tersendiri dalam mengolah dan menganalisis informasi tersebut secara manual. Oleh karena itu, diperlukan metode otomatisasi yang mampu mengklasifikasikan keterampilan kerja dengan akurat agar data tersebut dapat dimanfaatkan secara optimal.

Dalam hal pengolahan teks, metode TF-IDF (*Term Frequency-Inverse Document Frequency*) merupakan salah satu teknik yang banyak digunakan untuk mengekstraksi fitur penting dari kumpulan dokumen teks (Addiga & Bagui, 2022). TF-IDF mampu memberikan bobot pada kata-kata yang sering muncul di dalam suatu dokumen namun jarang muncul di dokumen lain, sehingga membantu mengidentifikasi kata kunci yang relevan dengan konten lowongan kerja (Nafis & awang, 2021). Dengan menggunakan TF-IDF, fitur-fitur yang merepresentasikan keterampilan kerja dapat diekstraksi secara efektif dari teks deskripsi lowongan kerja (Naeem et al., 2022).

Setelah fitur-fitur tersebut berhasil diekstraksi, tahap selanjutnya adalah melakukan klasifikasi keterampilan kerja berdasarkan fitur yang ada. *Decision Tree*

merupakan salah satu algoritma klasifikasi yang populer karena kemampuannya untuk menghasilkan model yang mudah dipahami dan diinterpretasikan (Costa & Pedreira, 2023). Algoritma ini juga cukup efektif dalam mengatasi data yang memiliki variabel kategorikal maupun numerik, sehingga cocok untuk digunakan dalam mengklasifikasikan jenis keterampilan kerja berdasarkan data teks yang telah diolah (Charbuty & Abdulazeez, 2021).

Penerapan metode *TF-IDF* dan *Decision T*ree dalam klasifikasi keterampilan kerja pada data lowongan LinkedIn dapat membantu mengotomatiskan proses identifikasi keterampilan yang dibutuhkan dalam berbagai bidang pekerjaan (Cheng et al., 2022). Hal ini tidak hanya mempercepat proses analisis data, tetapi juga meningkatkan akurasi dalam mengelompokkan keterampilan sehingga lebih sistematis dan informatif. Dengan klasifikasi yang baik, data keterampilan kerja menjadi lebih mudah diakses dan dimanfaatkan untuk berbagai kepentingan, mulai dari pencarian kerja, pelatihan keterampilan, hingga pengembangan kebijakan ketenagakerjaan (Singh & Tripathi, 2021). Dalam konteks ini, teknik text mining dapat dimanfaatkan untuk menggali informasi dari teks tidak terstruktur tersebut. Dengan menggunakan metode ini, kita dapat mengidentifikasi keterampilan (skills), jabatan (job titles), dan tren industri yang paling banyak diminati oleh perusahaan. Penelitian ini bertujuan untuk menganalisis tren keterampilan kerja berdasarkan data lowongan pekerjaan dari LinkedIn, sehingga hasilnya dapat dimanfaatkan oleh pencari kerja, institusi pendidikan, dan pembuat kebijakan.

Penelitian-penelitian terdahulu menunjukkan pemanfaatan yang luas dari metode *Text Mining* dan algoritma *Machine Learning* dalam pengolahan data berbasis teks, khususnya dalam konteks klasifikasi dan sistem rekomendasi. Penelitian oleh Fitria (2024) misalnya, menerapkan metode *TF-IDF* sebagai pembobotan kata dalam sistem rekomendasi kerja berbasis *content-based filtering*, menggunakan data dari LinkedIn dan JobStreet. Sistem tersebut menggunakan *cosine similarity* untuk mencocokkan profil pengguna dengan lowongan kerja. Hal ini menunjukkan bahwa *TF-IDF* sangat efektif dalam mengekstraksi fitur penting dari teks deskriptif, sehingga sangat relevan jika diterapkan pada data lowongan kerja yang kaya informasi tidak terstruktur.

Penelitian lain oleh Anggina et al. (2022) menggunakan kombinasi metode *Lexicon-Based* dan *TF-IDF* dalam analisis sentimen ulasan pelanggan menggunakan algoritma *Multinomial Naïve Bayes Classifier*. Hasil akurasi yang tinggi (95%) menunjukkan bahwa kombinasi teknik ekstraksi fitur berbasis teks dengan algoritma klasifikasi sederhana namun efektif dapat menghasilkan model yang akurat. Penelitian ini menyoroti pentingnya representasi teks yang baik untuk mendukung performa model. Sementara itu, penelitian oleh Kurniawan et al. (2025) membuktikan keunggulan *Random Forest* dalam konteks klasifikasi multi-label soal pelajaran, menunjukkan bahwa pemilihan algoritma juga sangat memengaruhi keberhasilan klasifikasi.

Penelitian yang dilakukan oleh Kurniawati (2024) dan Singgalen (2023) menunjukkan pemanfaatan algoritma *Decision Tree* dalam klasifikasi berbasis atribut dan teks. Kurniawati menggunakan *Decision Tree* untuk memprediksi

kelulusan mahasiswa berdasarkan IPK dan jalur masuk, menghasilkan akurasi tinggi (96,91%). Sementara itu, Singgalen menggunakan *Decision Tree* dengan teknik *SMOTE Upsampling* dalam klasifikasi sentimen wisatawan, mencapai akurasi hingga 98,27%. Hasil ini menunjukkan bahwa *Decision Tree* tidak hanya akurat tetapi juga mudah diinterpretasikan, sehingga sangat bermanfaat untuk analisis berbasis teks seperti deskripsi pekerjaan.

Berdasarkan bebrapa penelitian di atas dapat diketahui bahwa belum terdapat penelitian yang secara spesifik memanfaatkan gabungan metode *TF-IDF* dan *Decision Tree* untuk melakukan klasifikasi keterampilan kerja dari data deskripsi lowongan pekerjaan, khususnya di platform profesional seperti LinkedIn. Gap ini penting karena data lowongan pekerjaan umumnya memiliki struktur teks yang kompleks, dan keterampilan yang disebutkan sangat bervariasi tergantung pada industri dan posisi. Selain itu, sebagian besar penelitian terdahulu lebih berfokus pada klasifikasi sentimen, prediksi kelulusan, atau sistem rekomendasi berbasis kemiripan, bukan pada klasifikasi keterampilan kerja sebagai entitas terstruktur dari teks deskriptif.

Kebaruan dari penelitian ini terletak pada fokusnya untuk mengklasifikasikan keterampilan kerja secara otomatis dari teks deskripsi pekerjaan di LinkedIn menggunakan *TF-IDF* sebagai metode ekstraksi fitur dan *Decision Tree* sebagai algoritma klasifikasi. Ini menawarkan pendekatan yang sederhana namun transparan, berbeda dengan penelitian sebelumnya yang cenderung menggunakan algoritma kompleks seperti *Random Forest*, *XGBoost*, atau *SVM*. Selain itu, penelitian ini juga dapat memberikan kontribusi praktis dalam

sistem penyelarasan tenaga kerja (*job matching*) berbasis keterampilan, yang sangat dibutuhkan dalam pengembangan sistem rekrutmen otomatis atau perencanaan pelatihan kerja. Oleh karena itu, penelitian ini tidak hanya mengisi gap konseptual, tetapi juga memberikan nilai tambah dalam implementasi sistem informasi ketenagakerjaan berbasis kecerdasan buatan.

Penelitian ini penting dilakukan karena di era digital saat ini, pencarian pekerjaan dan pemetaan keterampilan yang dibutuhkan oleh pasar kerja semakin bergantung pada data online, seperti lowongan kerja yang dipublikasikan di platform LinkedIn. Dengan mengembangkan metode klasifikasi keterampilan kerja menggunakan TF-IDF dan Decision Tree, proses identifikasi keterampilan yang relevan dapat dilakukan secara otomatis, cepat, dan akurat. Hal ini sangat bermanfaat bagi pencari kerja, perusahaan, serta lembaga pelatihan dalam menyesuaikan kebutuhan kompetensi dengan tren pasar kerja terkini. Selain itu, penelitian ini juga dapat membantu dalam pengembangan sistem rekomendasi pekerjaan yang lebih efektif dan mendukung pengambilan keputusan yang berbasis data dalam dunia ketenagakerjaan.

1.2 Pernyataan Masalah

1. Bagaimana tren lowongan kerja berdasarkan waktu, lokasi, dan perusahaan dengan menggunakan teknik text mining untuk memahami perkembangan pasar kerja dari LinkedIn?
2. Bagaimana teknik pre-processing dan TF-IDF untuk mengolah teks secara sistematis dan mengekstrak informasi penting dari lowongan kerja?

3. Bagaimana hasil modeling klasifikasi menggunakan algoritma Decision Tree untuk mengelompokkan jenis pekerjaan berdasarkan deskripsi lowongan dan mengevaluasi kinerjanya menggunakan metrik evaluasi klasifikasi?

1.3 Tujuan Penelitian

1. Mengidentifikasi tren lowongan kerja berdasarkan waktu, lokasi, dan perusahaan dengan menggunakan teknik text mining untuk memahami perkembangan pasar kerja.
2. Menerapkan teknik pre-processing dan TF-IDF untuk mengolah teks secara sistematis dan mengekstrak informasi penting dari lowongan kerja.
3. Melakukan modeling klasifikasi menggunakan algoritma Decision Tree untuk mengelompokkan jenis pekerjaan berdasarkan deskripsi lowongan dan mengevaluasi kinerjanya menggunakan metrik evaluasi klasifikasi

1.4 Batasan Masalah

1. Data diambil hanya dari platform LinkedIn.
2. Fokus analisis pada teks judul dan deskripsi lowongan kerja.
3. Data yang digunakan dalam penelitian ini diambil dari API situs platform LinkedIn pada Mei 2025.

1.5 Manfaat Penelitian

1. Visualisasi data menggunakan dua metode yaitu Decision Tree dan TF-IDF

2. Menampilkan sub-kategori dan macam-macam skill informasi dan teknologi yang dihasilkan oleh pemodelan data mining

1.6 Sistematika Penulisan

Sistematika penulisan tesis terbagi menjadi 5 (lima) bab untuk memudahkan dalam penulisan, antara lain:

BAB I : PENDAHULUAN

Bab ini berisi latar belakang, rumusan masalah, tujuan, batasan masalah, manfaat penelitian, dan sistematika penulisan

BAB II : STUDI PUSTAKA

Bab ini berisi tentang teori-teori dasar terkait dengan penyusunan laporan penelitian yaitu, jobstreet, data mining, klasifikasi, TF-IDF dan decision tree dan penelitian terkait.

BAB III : METODOLOGI PENELITIAN

Bab ini berisi tentang waktu dan tempat penelitian, alat dan bahan penelitian, dan metode penelitian.

BAB IV : PEMBAHASAN

Bab ini berisi hasil dan pembahasan terkait dengan penyusunan laporan penelitian yaitu, pengumpulan dataset, preprocessing teks, analisis statistik deskriptif, penerapan teknik text mining, modeling klasifikasi pekerjaan dan analisis kebutuhan keterampilan. .

BAB V : KESIMPULAN

Bab ini memuat kesimpulan berdasarkan hasil pembahasan laporan penelitian, serta berisi saran perbaikan dan pengembangan lebih lanjut.

DAFTAR PUSTAKA

Bab ini memuat daftar sumber kutipan teori - teori yang dijadikan acuan dalam menulis laporan.

LAMPIRAN

Lampiran memuat dokumentasi berkas-berkas penunjang penulisan laporan, berupa foto, dan dokumen lain.

BAB II

STUDI PUSTAKA

2.1 Text Minning

Text mining (penambangan teks) adalah penambangan yang dilakukan oleh komputer untuk mendapatkan sesuatu yang baru, sesuatu yang tidak diketahui sebelumnya atau menemukan kembali informasi yang tersebar, yang berasal dari informasi yang diekstrak secara otomatis dari sumber-sumber data teks yang berbeda-beda *(Feldman & Sanger, 2007).* Text mining merupakan teknik yang digunakan untuk menangani masalah klasifikasi*, clustering, information extraction* dan *information retrival* (Berry & Kogan, 2010). Tahap-tahap text mining secara umum adalah *text preprocessing* dan *feature selection (Feldman & Sanger, 2007), (Berry & Kogan, 2010).*

Konsep text mining sangat relevan dengan prinsip ilmu pengetahuan dalam Islam yang menekankan pentingnya mencari dan menyingkap sesuatu yang tersembunyi. Dalam Al-Qur'an surat Thaha ayat 114, Allah SWT berfirman:

فَتَعَالَى اللَّهُ الْمَلِكُ الْحَقُّ ۗ وَلَا تَعْجَلْ بِالْقُرْآنِ مِنْ قَبْلِ أَنْ يُقْضَىٰ إِلَيْكَ وَحْيُهُ ۖ وَقُلْ رَبِّ زِدْنِي عِلْمًا

Artinya:

Maka Maha Tinggi Allah Raja Yang sebenar-benarnya, dan janganlah kamu tergesa-gesa membaca Al quran sebelum disempurnakan mewahyukannya kepadamu, dan katakanlah: "Ya Tuhanku, tambahkanlah kepadaku ilmu pengetahuan" (Q.S. Thaha: 114).

Tafsir Ringkas:

"وَلَا تَعْجَلْ بِالْقُرْآنِ مِنْ قَبْلِ أَنْ يُقْضَىٰ إِلَيْكَ وَحْيُهُ"

Artinya: *Janganlah kamu tergesa-gesa (membaca atau menyampaikan) Al-Qur'an sebelum wahyunya sempurna disampaikan kepadamu*.

Ini adalah peringatan kepada Nabi Muhammad ﷺ agar tidak terburu-buru saat menerima wahyu, agar beliau tidak salah dalam menyampaikan maknanya. Ini juga menunjukkan proses pewahyuan yang bertahap dan penuh ketelitian. (tafsir ibnu kathir, (1420H,2000)

"وَقُلْ رَبِّ زِدْنِي عِلْمًا"

Artinya: *Katakanlah, "Ya Tuhanku, tambahkanlah kepadaku ilmu*."

Ini menunjukkan keutamaan ilmu dan bahwa Nabi Muhammad ﷺ sendiri diperintahkan untuk terus memohon tambahan ilmu. Ini juga menjadi doa yang sangat penting dan dianjurkan untuk diamalkan oleh umat Islam. (tafsir ibnu kathir, 1420H,2000)

Ayat di atas mengandung pesan mendalam tentang etika dalam menuntut ilmu. Ayat ini menekankan dua hal penting, yaitu perlunya kesabaran dan kehati-hatian dalam menerima serta memahami informasi, serta semangat untuk terus menambah ilmu pengetahuan. Pesan ini sangat relevan dalam konteks perkembangan teknologi informasi dan ilmu data saat ini, salah

satunya dalam bidang text mining (penambangan teks). Hal ini terkait dengan hadis tentang bekerja dan mencari nafkah :

مَا أَكَلَ أَحَدٌ طَعَامًا قَطُّ خَيْرًا مِنْ أَنْ يَأْكُلَ مِنْ عَمَلِ يَدِهِ، وَإِنَّ نَبِيَّ اللَّهِ دَاوُدَ كَانَ يَأْكُلُ مِنْ عَمَلِ يَدِهِ

(رواه البخاري، رقم 2072)

Artinya:

"Tidaklah seseorang memakan makanan yang lebih baik dari hasil kerja tangannya sendiri. Dan sesungguhnya Nabi Allah Daud 'alaihis salam makan dari hasil kerja tangannya sendiri." (HR. Bukhari no. 2072)

Hadis ini Menekankan nilai pentingnya bekerja secara mandiri dan produktif, sesuai dengan semangat analisis keterampilan kerja dalam dunia profesional.

Hadis ini menekankan pentingnya bekerja dengan tangan sendiri dan mencari nafkah secara halal

1. Keutamaan Bekerja Sendiri

Rasulullah ﷺ menyatakan bahwa makanan terbaik adalah yang diperoleh dari hasil kerja tangan sendiri. Ini menunjukkan bahwa bekerja keras dan mandiri dalam mencari nafkah adalah tindakan yang mulia dan diberkahi.

2. Teladan dari Nabi Dawud عليه السلام

Nabi Dawud عليه السلام, meskipun sebagai seorang raja dan nabi, memilih untuk makan dari hasil kerja tangannya sendiri. Ini memberikan contoh bahwa tidak ada pekerjaan yang hina selama dilakukan dengan cara yang halal dan jujur.

3. Anjuran untuk Mandiri dan Tidak Bergantung pada Orang Lain Hadis ini mendorong umat Islam untuk tidak bergantung pada orang lain dalam memenuhi kebutuhan hidupnya. Sebaliknya, mereka dianjurkan untuk berusaha sendiri dan tidak meminta-minta.(Ibn Hajar al-'Asqalani p. 306)

Surah Al-Mulk ayat 1 dalam Al-Qur'an

تَبٰرَكَ الَّذِيْ بِيَدِهِ الْمُلْكُۖ وَهُوَ عَلٰى كُلِّ شَيْءٍ قَدِيْرُۙ

Artinya: Mahasuci Allah yang menguasai (segala) kerajaan, dan Dia Mahakuasa atas segala sesuatu.

Ibnu Katsir menjelaskan bahwa ayat ini mengagungkan Dzat Allah yang memiliki kekuasaan penuh atas kerajaan alam semesta. Segala sesuatu berada di bawah pengaturan dan kehendak-Nya, tiada yang menggugurkan qadar-Nya—karena kebijaksanaan, keadilan, dan kekuasaan-Nya(tafsir ibnu kathir, 1420H,2000)

أَلَّذِي خَلَقَ الْمَوْتَ وَالْحَيَاةَ لِيَبْلُوَكُمْ أَيُّكُمْ أَحْسَنُ عَمَلًا ۚ وَهُوَ الْعَزِيزُ الْغَفُورُ

Artinya:

Yang menciptakan mati dan hidup, untuk menguji kamu siapa di antara kamu yang lebih baik amalnya. Dan Dia Maha Perkasa lagi Maha Pengampun.

Ibnu Katsir menerangkan bahwa penciptaan kematian dan kehidupan ditujukan sebagai sarana ujian terhadap manusia, untuk menyeleksi siapa yang lebih baik amalnya, bukan sekadar yang paling banyak.

Gelarnya sebagai Al-Aziz (Maha Perkasa) dan Al-Ghafur (Maha Pengampun) menegaskan bahwa meskipun Ia memiliki kekuasaan mutlak, namun kesalahan dan dosa tetap diberi kesempatan untuk diampuni bagi yang bertaubat
(tafsir ibnu kathir, (1420H,2000)

Meskipun ayat ini berbicara tentang hikmah penciptaan dan ujian dalam kehidupan, nilai-nilainya memiliki makna konseptual yang sangat relevan dalam dunia teknologi informasi. Setiap perangkat lunak *(software*) harus diuji sebelum dirilis, seperti dengan unit test, integration test, atau load testing. Tujuannya bukan hanya mencari yang berfungsi, tetapi yang paling efisien, aman, dan optimal → *"aḥsanu 'amalā"* (amal terbaik). *Nilai spiritual ini memberi dimensi moral dan filosofis dalam dunia TI*, agar manusia tidak hanya menciptakan teknologi yang hebat, tetapi juga bermakna dan bertanggung jawab.

Text mining merupakan proses sistematis yang dilakukan untuk menggali informasi dan pengetahuan tersembunyi dari data teks dalam jumlah besar. Dalam pelaksanaannya, text mining membutuhkan ketelitian dalam tahap preprocessing, pemilihan fitur (*feature selection*), serta dalam penerapan algoritma analisis seperti klasifikasi atau klasterisasi. Hal ini mencerminkan pesan dalam ayat tersebut agar tidak tergesa-gesa dalam membaca atau

memahami teks, melainkan harus melalui proses yang terstruktur dan mendalam. Selain itu, upaya dalam text mining untuk menemukan wawasan baru dari data teks mencerminkan permohonan yang tersirat dalam doa, “Ya Tuhanku, tambahkanlah kepadaku ilmu pengetahuan.” Artinya, teknologi text mining bisa dilihat sebagai bentuk nyata dari ikhtiar manusia untuk menambah ilmu melalui analisis data yang tidak mungkin dilakukan secara manual.

Dengan demikian, ayat ini tidak hanya menjadi pedoman spiritual, tetapi juga menjadi dasar filosofis dalam mendukung aktivitas ilmiah modern seperti text mining. Teknologi ini menjadi sarana untuk memahami informasi yang tersebar dan tersembunyi, sehingga membantu manusia dalam menyingkap pengetahuan baru, sejalan dengan semangat Islam yang mendorong pencarian ilmu secara terus-menerus dan penuh kehati-hatian..

2.2 TF-IDF (*Term Frequency - Inverse Document Frequency*)

Term Frequency - Inverse Document Frequency (TF-IDF) adalah sebuah metode yang berfungsi untuk memberikan nilai bobot untuk teks atau kata yang terdapat didalam kalimat. TF-IDF merupakan metode yang menggabungkan dua konsep dalam perhitungannya, yaitu *term frequency* dan *inversed document frequeny. Term Frequency* adalah frekuensi kata dalam dokumen atau kalimat sedangkan, *inversed document frequency* adalah frekuensi dari dokumen yang mengandung kata tersebut. Frekuensi dalam kemunculan suatu kata diberikan untuk menunjukan seberapa penting kata (Ramos, 2003). Rumus dalam melakukan tahap TF-IDF dapat dilihat pada rumus 2.1, 2.2 dan 2.3.

2.2.1 Rumus perhitungan TF

Term Frequency (TF) merupakan pembobotan kata dalam dokumen dengan rumus sebgai berikut:

$$tf(t.d) = \frac{tf}{\max(tf)} \quad \text{..... (2.1)}$$

Keterangan:

tf(t,d) = Frekuensi yang terdapat pada term (TF).

max(tf) = Total dari keseluruhan kata yang terdapat dalam dokumen.

tf = Jumlah dari kemunculan term terbanyak. didalam dokumen yang sama.

2.2.2 Rumus perhitungan IDF

Inversed document frequency (IDF) adalah frekuensi dari dokumen yang mengandung kata dengan rumus sebagai berikut:

$$idft = \log(\frac{D}{df_t}) \quad \text{..... (2.2)}$$

Keterangan:

D = Total kesuluruhan dari dokumen.

idf(t) = bobot pada kemunculan term t pada setiap dokumen.

d ft = jumlah dari dokumen yang mengandung term t

2.2.3 Rumus perhitungan TF-IDF

$W_{t,d} = tf(t,d)x\ idf_t$..... (2.3)

Keterangan:

Wt ,d = bobot term dalam suatu dokumen

idf(t) = bobot dari kemunculan term t pada setiap dokumen.

tf(t,d) = Frekuensi term (TF)

2.3 Decision Tree

Decision tree method adalah alat statistik yang kuat untuk klasifikasi, prediksi, interpretasi, dan manipulasi data yang memiliki beberapa aplikasi potensial dalam penelitian (Tafseer,2018). *Decision tree models* biasanya digunakan untuk berbagai tugas, termasuk :

1. *Variable selection*

   Seiring dengan semakin lazimnya penyimpanan data terkomputerisasi, jumlah variabel yang dilacak dalam pengaturan klinis telah meningkat secara signifikan. Banyak dari variabel-variabel ini hanya sesekali relevan, sehingga lebih baik untuk menghilangkannya dari aktivitas penggalian data. Metode decision tree dapat mengidentifikasi variabel input yang paling relevan untuk digunakan dalam model pohon keputusan, membantu perumusan hipotesis klinis dan mengarahkan penelitian selanjutnya. Pendekatan ini mirip dengan pemilihan variabel bertahap yang ditemukan dalam analisis regresi.

2. *Assessing the relative importance of variables*

   Setelah mengidentifikasi sekumpulan variabel yang relevan, para peneliti sering kali berusaha untuk memahami peran penting yang dimainkan oleh variabel-variabel ini. Tingkat kepentingan variabel biasanya ditentukan dengan menilai seberapa besar akurasi model (atau kemurnian simpul dalam tree) berkurang ketika sebuah variabel dihilangkan. Secara umum, semakin besar dampak dari sebuah variabel terhadap jumlah record yang lebih besar, maka variabel tersebut dianggap semakin penting.

3. *Handling of missing values*

   Menangani data yang hilang dengan mengecualikan kasus-kasus dengan nilai yang hilang adalah pendekatan yang umum dilakukan namun tidak tepat. Metode ini tidak efisien dan dapat menimbulkan bias dalam analisis. Analisis decision tree menawarkan dua pendekatan alternatif untuk menangani data yang hilang. Pertama, ia dapat memperlakukan nilai yang hilang sebagai kategori yang terpisah, yang memungkinkan mereka untuk dianalisis bersama kategori lainnya. Atau, decision tree model dapat dibangun dengan menggunakan variabel dengan banyak nilai yang hilang sebagai variabel target untuk prediksi, dan mengganti nilai yang hilang tersebut dengan hasil prediksi.

4. *Prediction*

   Salah satu aplikasi utama dari decision tree model adalah memprediksi hasil di masa depan. Dengan memanfaatkan data historis dan membangun tree model, para peneliti dapat dengan mudah meramalkan hasil untuk catatan yang akan datang.

5. *Data manipulation*

   Dalam penelitian, adalah hal yang umum untuk menemukan terlalu banyak kategori untuk satu variabel kategorikal atau data kontinu yang sangat miring. Decision tree model dapat membantu dengan menyarankan cara menyederhanakan variabel kategorikal menjadi jumlah kategori yang lebih mudah dikelola atau cara membagi variabel yang condong ke dalam rentang variabel.

2.4 LinkedIn

LinkedIn adalah jaringan profesional terbesar di dunia dengan lebih dari 120 juta anggota dan berkembang pesatt yang menghubungkan pengguna ke kontak tepercaya dan membantu pengguna untuk bertukar pengetahuan, ide, dan peluang dengan jaringan profesional yang lebih luas (Case, 2013). LinkedIn adalah jejaring sosial berbasis internet untuk pencarian kerja dan posting pekerjaan, dan jaringan serta kemampuan untuk menemukan kandidat unik adalah alasan utama untuk berada di platform ini (Larsen, 2020). LinkedIn dapat digunakan dalam bentuk berbasis web di komputer, atau dengan aplikasi di smartphone. Di LinkedIn, pengguna memiliki kesempatan untuk membuat profil

di mana mereka dapat membuat daftar karakteristik mereka yang paling penting, pendidikan yang relevan, pengalaman dan pekerjaan masa lalu dan pekerjaan saat ini. koneksi membuka kesempatan bagi pengguna untuk mengikuti organisasi, orang-orang yang dikenal orang tersebut dan orang-orang yang ingin dihubungi atau dianggap menarik.

LinkedIn dapat digunakan sebagai platform untuk membentuk penjemanaam pribadi atau personal branding guna memberikan impresi dan citra yang positif yang tertanam di jejaring media sosial. Pembentukan penjenamaan pribadi ini juga dilakukan oleh mahasiswa guna saling terhubung dengan pengguna lain yang memiliki interest atau ketertarikan yang sama, seperti antara mahasiswa dengan dosen, mahasiswa dengan recruiter, jobseeker dengan recruiter, dan jaringan profesional lainnya yang sejenis. Dalam penjenamaan pribadi di LinkedIn, mahasiswa dapat menunjukkan kapabilitasnya di bidang tertentu yang dapat diunggah di laman LinkedIn dalam bentuk konten guna menjadi daya tarik tersendiri bagi perekrut (Salin, 2017). Dengan menggunakan LinkedIn juga, mahasiswa atau setiap pengguna LinkedIn dapat berinteraksi dengan pengguna lainnya guna berbagi informasi mengenai portofolio berupa pengalaman kerja, magang, atau projek tugas perkuliahan yang dapat “menjual” dan menambah value dari pengguna LinkedIn tersebut (Ayu et al., 2020) Dengan kapabilitas, value, dan portofolio yang dimiliki serta diunggah di LinkedIn, yang mana tendensi dan fokus LinkedIn ini berada di ranah profesional, membuat mahasiswa, khususnya mahasiswa tingkat akhir, setidaknya bisa melangkah selangkah lebih maju dalam mempersiapkan karir pascakelulusan. LinkedIn

sebagai media sosial yang berfokus terhadap ranah profesional memberikan kemudahan bagi setiap pengguna – mahasiswa guna mengunggah segala pencapaian profesional yang diraih di laman LinkedIn sehingga terdapat sebuah paradgima di tengah-tengah mahasiswa bahwasannya pengguaan LinkedIn di kalangan mahasiswa menjadi sebuah ajang untuk menunjukkan identitas diri secara profesional (Ayu et al., 2020).

Chandler (dalam Ayu et al., 2020) dalam membangun indetitas virtual, homepage adalah tempat bagi seseorang untuk menggungah konten pribadi dengan tujuan tertetu, sehingga sebagai media sosial berbasir profesional, LinkedIn memberikan ruang bagi mahasiswa untuk membangun realitas virtual karena dapat dipergunakan sedemikian rupa oleh penggunannya yang mana identitas virtual tersebut direpresentasikan dari setiap unggahan konten pengguna LinkedIn yang telah tertata profilnya dengan mencantumkan beberapa informasi detail mengenai latar belakang pendidikan, interest, pengalaman kerja/magang, dan sebagainya.

Metode *decision tree* adalah salah satu alat statistik yang digunakan secara luas dalam klasifikasi dan prediksi data. Metode ini memiliki kekuatan dalam membantu proses pengambilan keputusan melalui struktur pohon bercabang yang menyerupai alur logika manusia. Setiap cabang menggambarkan pilihan-pilihan berdasarkan atribut data, dan hasil akhirnya mengarahkan pada sebuah keputusan yang logis dan terukur. Model ini banyak diterapkan dalam berbagai penelitian karena kemampuannya untuk menafsirkan data yang kompleks menjadi bentuk yang mudah dipahami (Solehuddin et al., 2022).

Dalam perspektif Islam, konsep pengambilan keputusan yang rasional, bertahap, dan berdasarkan pertimbangan yang matang juga ditekankan dalam Al-Qur'an. Salah satu ayat yang dapat dikaitkan dengan prinsip di balik *decision tree* adalah firman Allah SWT dalam Qur'an.Surah. Az-Zumar ayat 18 sebagai berikut:

ٱلَّذِينَ يَسْتَمِعُونَ ٱلْقَوْلَ فَيَتَّبِعُونَ أَحْسَنَهُۥٓ ۚ أُو۟لَٰٓئِكَ ٱلَّذِينَ هَدَىٰهُمُ ٱللَّهُ
وَأُو۟لَٰٓئِكَ هُمْ أُو۟لُوا۟ ٱلْأَلْبَٰبِ

Artinya:

Yang mendengarkan perkataan lalu mengikuti apa yang paling baik di antaranya. Mereka itulah orang-orang yang telah diberi Allah petunjuk dan mereka itulah orang-orang yang mempunyai akal.

Tafsir Ibnu Katsir menjelaskan bahwa ayat ini memuji orang-orang yang:

1. Mendengarkan berbagai bentuk perkataan dan ajaran, termasuk yang baik maupun buruk.
2. Memilih dan mengikuti perkataan yang terbaik, yaitu wahyu Allah, kebenaran, dan ajaran para rasul.

Perkataan yang terbaik ini bisa berarti:

1. Al-Qur'an, yang merupakan perkataan paling sempurna.
2. Ucapan yang mengajak kepada kebaikan, seperti dakwah, nasihat, dan ilmu yang bermanfaat.

Mereka diberi sifat "ulul albab" (orang-orang berakal), karena mereka memiliki kemampuan membedakan mana yang benar dan terbaik untuk diikuti.

Ibnu Katsir juga menyebut bahwa ayat ini mencakup semua orang yang mau berpikir, bersikap kritis, dan jujur dalam menilai kebenaran (tafsir ibnu kathir, (1420H,2000)

2.5 *Preprocessing* Data Teks

Text-preprocessing merupakan tahapan yang memiliki fungsi untuk menyeleksi teks menjadi terstruktur melalui serangkaian tahapan (Sari,2020). Dalam penelitian ini *ext-prepprocessing* memiliki beberapa tahapan yaitu;

1. *Case Folding*

   Case Folding merupakan tahapan dalam *text-preprocessing* yang memiliki fungsi untuk mengubah teks menjadi huruf kecil atau *lowercase.*

2. *Cleaning Data*

   Menyeleksi dan membersihkan data teks dari *symbol, emoticon, link, single char, number, whites pace* dan *punctuation* yang tidak memiliki sentimen dalam suatu teks yang terdapat dalam kalimat atau dokumen.

3. *Tokenizing*

   Tahapan *Tokenizing* merupakan tahapan dalam *text-preprocessing* yang memiliki fungsi sebagai pemisah kalimat pada data menjadi daftar kata. Tahapan ini dilakukan karena untuk mengimplementasikan tahapan selanjutnya dalam *text-preprocessing* yaitu; *Normalisasi, Stopwords* dan *Stemming*.

4. *Normalisasi*

   Normalisasi berfungsi untuk memperbaiki salah penulisan kata dalam teks atau *typo*. Contoh dari penggunaan normalisasi yaitu kata ”mrpkn” menjadi kata ”merupakan”.

5. *Stopwords*

   Stopwords adalah tahapan dalam *text-preprocessing* yang memiliki fungsi untuk memilih kata-kata yang dianggap penting dalam sebuah teks dan menghilangkan kata-kata dan simbol yang tidak bermakna dalam teks seperti konjungsi.

6. *Stemming*

   Stemming merupakan tahapan akhir dari melakukan tahapan textpreprocessing yang memiliki fungsi untuk mengubah kata yang memiliki imbuhan menjadi bentuk kata dasar. Contoh penggunaan stemming dalam teks adalah kata ”melakukan” dirubah menjadi kata dasarnya yaitu ”lakukan”.

2.6 Penelitian Terdahulu

Tabel 2. 1 Penelitian Terdahulu

No	Nama, Tahun,, Judul	Metode	Hasil
1	Fitria, A. (2024), Sistem Rekomendasi Pekerjaan Menggunakan Pendeketan Content-Based Filtering	Penelitian ini menggunakan metode *content-based filtering* untuk sistem rekomendasi pekerjaan. Data diperoleh melalui *web scraping* dari JobStreet (lowongan) dan LinkedIn (profil pencari kerja). Analisis dilakukan dengan *text preprocessing*, pembobotan TF-IDF, dan perhitungan *cosine similarity*.	Sistem rekomendasi memberikan hasil presisi rata-rata 0,53 dan lulus uji fungsional dengan *blackbox testing*. Penelitian menyimpulkan bahwa pendekatan ini efektif dalam mencocokkan profil pencari kerja dengan lowongan yang relevan.
2	Singgalen (2023), Penerapan Metode CRISP-DM dalam Klasifikasi Data Ulasan Pengunjung Destinasi Danau Toba Menggunakan Algoritma Naïve Bayes Classifier (NBC) dan Decision Tree (DT)	Menggunakan tahapan CRISP-DM dan algoritma NBC serta DT untuk klasifikasi data ulasan dari Tripadvisor. Data terdiri dari 858 ulasan dan diproses menggunakan SMOTE untuk penyeimbangan data.	DT dengan SMOTE memberikan performa tertinggi: akurasi 98,27%, presisi 98,83%, recall 97,71%, f-measure 98,26%, AUC 0,982. Penelitian menunjukkan pentingnya pelayanan prima dan infrastruktur dalam persepsi wisatawan.
3	Rahayu, Fauzi& Indra (2022), Analisis Sentimen Terhadap Program Kampus Merdeka Menggunakan Naive Bayes Dan Support Vector Machine	Menggunakan data Twitter bertagar #kampusmerdeka (2019–2022), dilakukan analisis sentimen dengan Naïve Bayes dan SVM. Data sebanyak 1118, terdiri dari 618 sentimen positif dan 500 negatif.	Hasil evaluasi menunjukkan bahwa Naïve Bayes mencapai akurasi sebesar 86%, presisi 87%, dan recall 80%, sementara SVM memberikan performa yang lebih baik dengan akurasi 93%, presisi 100%, dan recall 84%. Dengan demikian, SVM terbukti lebih akurat dalam mengklasifikasikan sentimen terhadap kebijakan Merdeka Belajar Kampus Merdeka (MBKM).
4	**Kusnaya, Cahyana, & Juwita (2025),**	Dataset berjumlah 627 tweet, dibagi 80% latih	Hasilnya menunjukkan bahwa Naïve Bayes

	Penerapan Metode Naive Bayes Multinomial dan Complement dalam Membandingkan Tingkat Akurasi terhadap Analisis Sentimen Kurikulum Merdeka	dan 20% uji. Digunakan algoritma Naïve Bayes Multinomial dan Complement dengan evaluasi menggunakan *confusion matrix*.	Multinomial menghasilkan akurasi sebesar 89%, sedikit lebih unggul dibandingkan algoritma Complement yang mencapai akurasi 88%. Penelitian ini menyimpulkan bahwa Naïve Bayes Multinomial lebih efektif dalam menganalisis sentimen terkait Kurikulum Merdeka dibandingkan metode Complement.
5	Anggina, Setiawan, & Bachtiar (2022),Analisis Ulasan Pelanggan Menggunakan Multinomial Naïve Bayes Classifier dengan Lexicon-Based dan TF-IDF Pada Formaggio Coffee and Resto.	Pengumpulan data melalui web scraping dari situs ulasan (Traveloka, Zomato, dll) sebanyak 741 ulasan (2018–2021).Analisis sentimen menggunakan Lexicon-Based, TF-IDF, dan algoritma **Multinomial Naïve Bayes Classifier**.Evaluasi model menggunakan **confusion matrix**: akurasi, presisi, recall, f1-score.	Evaluasi model menggunakan confusion matrix menghasilkan akurasi sebesar 95%, presisi 85%, recall 68%, dan F1-score 72%. Hasil analisis ini divisualisasikan dalam sebuah dashboard yang memperoleh skor Sistem Usability Scale (SUS) sebesar 67,5, yang menunjukkan penerimaan yang baik dari pihak manajemen Formaggio.
6	Kurniawan, Pratiwi, & Hamami. (2025), asifikasi Soal Menggunakan Multi-Label Problem Transformation dengan Metode Random Forest dan K-Nearest Neighbor.	klasifikasi soal Bahasa Indonesia tingkat SMP menggunakan metode multi-label classification dengan algoritma Random Forest dan K-Nearest Neighbor (K-NN). Pendekatan problem transformation yang digunakan meliputi Binary Relevance, Classifier Chain, dan Label Powerset, dan evaluasi model dilakukan melalui F1-	Hasil menunjukkan bahwa Random Forest dengan metode Label Powerset memberikan performa terbaik dengan F1-score sebesar 69%, sedangkan K-NN hanya mencapai 44%. Penelitian ini menyimpulkan bahwa Random Forest lebih efektif untuk klasifikasi multi-label soal dalam sistem Learning Management System (LMS).

		score serta K-Fold Cross Validation.	
7	Kurniawati (2024), Klasifikasi Data Mahasiswa Lampau Menggunakan Metode Decision Tree dan Support Vector Machine.	menggunakan data mahasiswa angkatan 2018 di UIN Maulana Malik Ibrahim dengan atribut IPK dan jalur masuk untuk memprediksi kelulusan menggunakan algoritma Decision Tree (DT) dan Support Vector Machine (SVM). Implementasi dilakukan dengan bahasa pemrograman Python dan library scikit-learn.	Dari hasil evaluasi, Decision Tree mencapai akurasi sebesar 96,91% dan SVM sebesar 96,62%, sehingga Decision Tree dinilai lebih unggul. Penelitian ini memiliki keunikan pada penggunaan kombinasi atribut akademik untuk prediksi kelulusan serta perbandingan kinerja dua algoritma machine learning tersebut.
8	Arif (2023), Perbandingan Kinerja Algoritma Random Forest, XGBoost dan LightGBM dalam Klasifikasi Emosi Komentar Reddit.	membandingkan kinerja algoritma Random Forest, XGBoost, dan LightGBM dalam klasifikasi emosi komentar pada platform Reddit menggunakan dataset GoEmotions Fine-Grained yang terdiri dari 7.325 komentar dengan 5 kategori emosi dasar. Fitur diekstraksi menggunakan CountVectorizer dan TF-IDF, sementara evaluasi model melibatkan Cross Validation, confusion matrix, dan hyperparameter tuning dengan GridSearchCV.	Hasil penelitian menunjukkan bahwa Random Forest menghasilkan akurasi tertinggi sebesar 75,38%, diikuti oleh XGBoost 69,05% dan LightGBM 66,63%. Kesimpulannya, Random Forest paling efektif untuk klasifikasi emosi pada komentar Reddit.

BAB III

METODOLOGI PENELITIAN

3.1 Jenis Penelitian

Penelitian ini merupakan penelitian kuantitatif deskriptif dan eksploratif, dengan pendekatan text mining. Tujuan utamanya adalah untuk menganalisis tren kebutuhan tenaga kerja di sektor Teknologi Informasi (IT) berdasarkan data lowongan kerja. Penelitian ini menggabungkan metode TF-IDF untuk ekstraksi fitur teks, algoritma decision tree untuk klasifikasi jenis pekerjaan, serta visualisasi berbasis wordcloud untuk memperkuat interpretasi hasil.

3.2 Sumber dan Jenis Data

Data yang digunakan adalah data simulasi lowongan kerja yang disusun secara manual dengan struktur yang menyerupai data dari situs LinkedIn. Dataset terdiri dari 100 entri lowongan kerja sektor IT, dengan atribut sebagai berikut:

1. Judul pekerjaan
2. Nama perusahaan
3. Lokasi
4. Tanggal posting
5. Deskripsi pekerjaan
6. Industri

Data deskripsi pekerjaan berisi informasi mengenai tanggung jawab dan keterampilan yang dibutuhkan, ditulis dalam bahasa Indonesia dan Inggris secara campuran, mencerminkan kondisi lowongan kerja sebenarnya.

3.3 Teknik Pengumpulan Data

Karena keterbatasan akses terhadap API LinkedIn, data dikumpulkan dengan cara simulasi (*synthetic data generation*) menggunakan pustaka *faker* di Python. Data ini dikurasi agar tetap menyerupai format lowongan kerja yang umum di platform profesional.

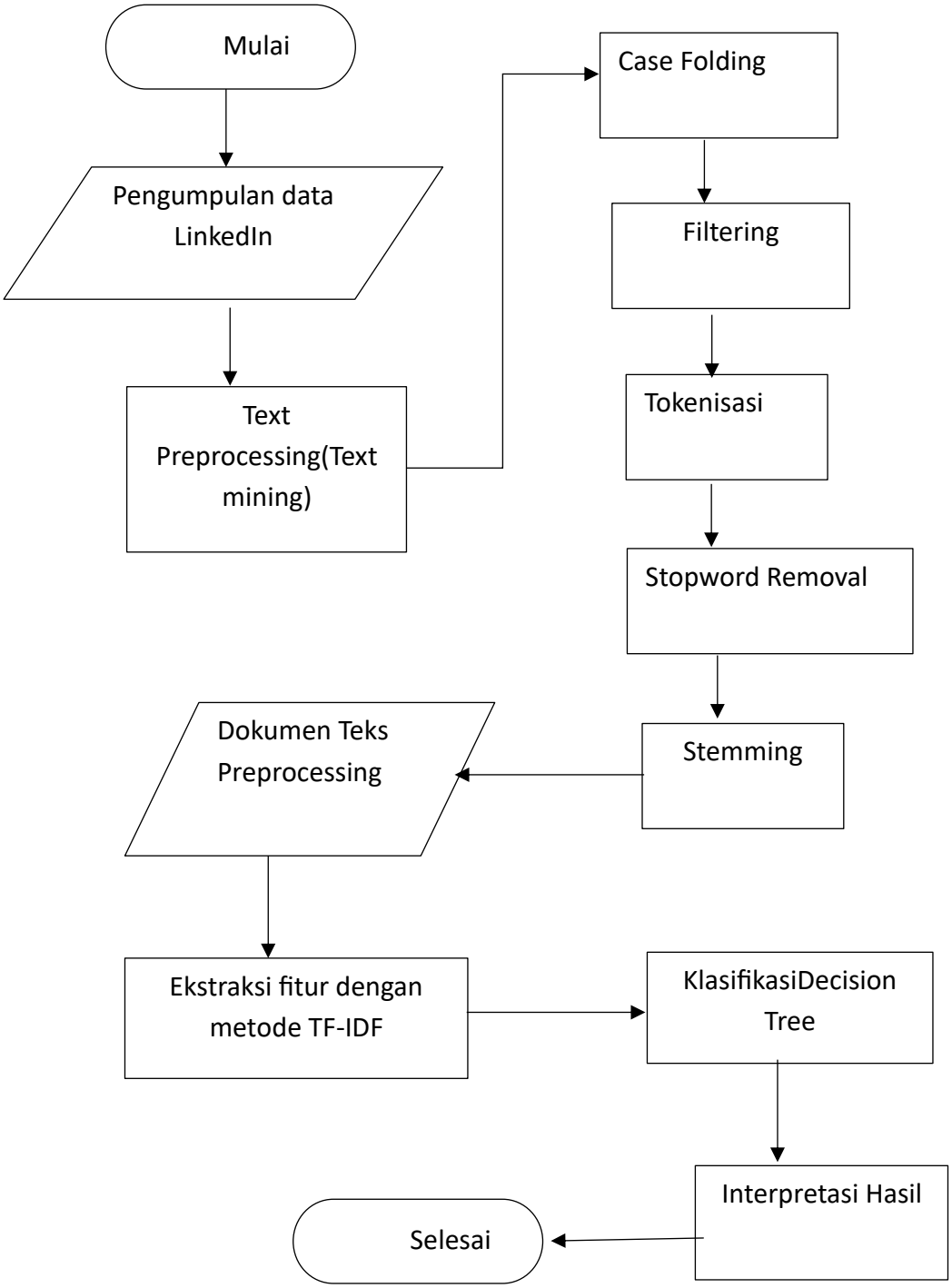
3.4 Alat dan Perangkat Lunak

Penelitian ini menggunakan sejumlah alat dan pustaka Python, yaitu:

- Python 3.x
- Pandas – manipulasi data
- Scikit-learn – TF-IDF dan decision tree
- Matplotlib & Seaborn – visualisasi data numerik
- WordCloud – visualisasi kata kunci
- Regex (re) – pembersihan teks

3.5 Tahapan Penelitian

Penelitian dilakukan melalui beberapa tahapan sebagai berikut:



Gambar 3. 1 Diagram Alur Penelitian

3.5.1 *Preprocessing* Teks

Tahapan ini bertujuan untuk membersihkan data teks deskripsi pekerjaan. Proses yang dilakukan meliputi:

1. Konversi teks menjadi huruf kecil
2. Penghapusan angka dan tanda baca
3. Penghapusan *stopwords* (kata umum) menggunakan daftar Bahasa Indonesia
4. Normalisasi teks

3.5.2 Ekstraksi Fitur Teks dengan TF-IDF

Setelah preprocessing, dilakukan ekstraksi fitur menggunakan TF-IDF (Term Frequency–Inverse Document Frequency). Teknik ini mengukur seberapa penting suatu kata dalam dokumen relatif terhadap keseluruhan dokumen. Hasil ekstraksi ini digunakan baik untuk analisis tren maupun untuk keperluan klasifikasi.

3.5.3 Klasifikasi Menggunakan Decision Tree

Untuk mengklasifikasikan jenis pekerjaan berdasarkan deskripsi, digunakan algoritma Decision Tree Classifier dari pustaka scikit-learn. Label klasifikasi yang digunakan adalah kolom judul_pekerjaan yang telah dikategorikan.

Langkah-langkah:

1. Melatih model dengan data TF-IDF sebagai input
2. Menggunakan 80% data untuk pelatihan dan 20% untuk pengujian
3. Mengukur akurasi model untuk menilai performa klasifikasi

3.5.4 Visualisasi WordCloud

Untuk memperjelas hasil analisis tren kata kunci, digunakan visualisasi wordcloud berdasarkan gabungan seluruh deskripsi pekerjaan. Kata-kata yang lebih sering muncul akan ditampilkan lebih besar, sehingga dapat terlihat keterampilan atau teknologi apa yang paling dominan dibutuhkan.

3.5.5 Teknik Analisis Data

Analisis dilakukan dalam dua pendekatan utama yaitu TF-IDF dan decision tree.

3.5.5.1 TF-IDF

TF-IDF (*Term Frequency Inverse Document Frequency*) merupakan metode yang digunakan untuk menentukan nilai frekuensi sebuah kata di dalam sebuah dokumen atau artikel dan juga frekuensi di dalam banyak dokumen. Perhitungan ini menentukan seberapa relevan sebuah kata di dalam sebuah dokumen (Evan, 2014). TFIDF adalah sebuah algoritma yang umumnya digunakan untuk pengolahan data besar (Kamath, 2014). Algoritma TF-IDF melakukan pemberian bobot pada setiap kata kunci disetiap kategori untuk mencari kemiripan kata kunci dengan kategori yang tersedia. Sebelum melakukan pembobotan maka akan dilakukan lima tahap pencarian text preprocessing yaitu pemecahan kalimat, case folding,

tokenizing, filtering, dan stemming, lalu selanjutnya dilakukan proses menghitung bobot TF-IDF, bobot query relevance dan bobot similarity (Marlinda & Rianto, 2013). Berdasarkan penelitian-penelitian sebelumnya, yang membahas tentang penerapan metode TF-IDF. Penulis menemukan banyak terdapat variasi formula dalam mengimplementasikan metode TF-IDF pada pembobotan kata. Nilai TF-IDF meningkat secara proporsional berdasarkan jumlah atau banyaknya kata yang muncul pada dokumen, tetapi diimbangi dengan frekuensi kata dalam korpus. Variasi dari skema pembobotan TF-IDF sering digunakan oleh mesin pencari sebagai alat utama dalam mencetak nilai (scoring) dan peringkat (ranking) sebuah relevansi dokumen yang diberikan user. TF-IDF pada dasarnya merupakan hasil dari perhitungan antara TF (Term Frequency) dan IDF (Inverse Document Frequency). Banyak cara untuk menentukan nilai yang tepat dari kedua statistik yang ada. Dalam kasus term frequency tf (t, d), cara yang paling sederhana adalah dengan menggunakan raw frequency di dalam dokumen, yaitu berapa kali term t muncul di dokumen d. Jika menyatakan raw frequency t sebagai f (t,d), maka skema tf yang sederhana adalah tf (t, d) = f (t,d). Nilai idf sebuah term (kata) dapat dihitung menggunakan persamaan sebagai berikut:

$$IDF = \log(\frac{D}{dfi})$$

D adalah jumlah dokumen yang berisi term (t) dan dfi adalah jumlah kemunculan (frekuensi) kata terhadap D. Adapun algoritma yang digunakan

untuk menghitung bobot (W) masingmasing dokumen terhadap kata kunci (query), yaitu :

$$Wd,t = tf\ d,t * IDFt$$

Keterangan :

d = dokumen ke–d

t = kata ke–t dari kata kunci

W = bobot dokumen ke–d terhadap kata ke–t

tf = term frekuensi/frekuensi kata

Setelah bobot (W) masing-masing dokumen diketahui, maka dilakukan proses pengurutan (sorting) dimana semakin besar nilai W, semakin besar tingkat kesamaan (similarity) dokumen tersebut terhadap kata yang dicari, demikian pula sebaliknya.

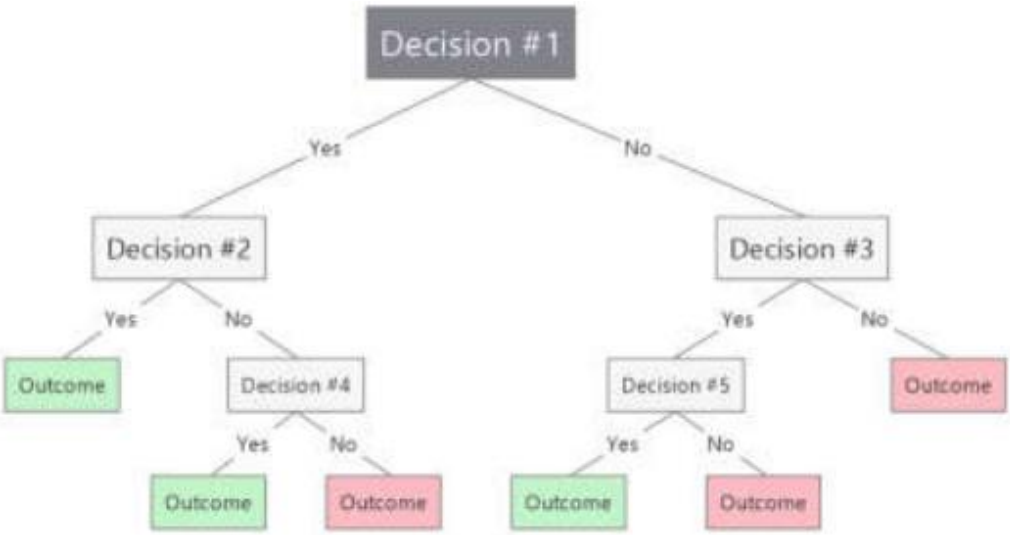
3.5.5.2 Decision Tree

Decision tree merupakan metode klarifikasi yang umum karena mudah diinterpretasikan oleh manusia. Decision Tree digunakan untuk mengklasifikasikan data, memprediksi pola dari data, dan menggambarkan hubungan antara variabel atribut x dan variabel target y dalam bentuk pohon atau struktur berhierarki. Decision Tree melakukan klasifikasi data dengan aturan yang telah ditetapkan untuk setiap atribut data. Dengan menggunakan Decision Tree proses pengambilan keputusan, klasifikasi ataupun prediksi yang awalnya kompleks akan menjadi lebih sederhana

karena memecah proses-proses yang ada menjadi lebih kecil. Kelebihan menggunakan Decision Tree antara lain:

1. Lebih mudah dipahami dan dianalisis
2. Dapat digunakan untuk mengeliminasi data yang tidak diperlukan
3. Lebih mudah untuk dibuat menjadi sebuah kesimpulan
4. Menangani kumpulan data non-linier secara efektif

Secara umum, klasifikasi Decision Tree dapat digambarkan sebagai berikut:



Gambar 3. 2 Contoh Decision Tree

3. 7 Kriteria Keberhasilan

Keberhasilan penelitian ini diukur dari:

1. Kemampuan model TF-IDF + decision tree dalam mengklasifikasikan judul pekerjaan secara akurat

2. Diperolehnya kata-kata kunci yang relevan dengan tren teknologi dan keterampilan di industri IT
3. Visualisasi wordcloud yang memperkuat temuan tekstur
4. Mendapatkan medel klasifikasi dengan metode DT
5. Evaluasi model DT

BAB IV

PEMBAHASAN

4. 1 Gambaran Umum Dataset

Penelitian ini menggunakan dataset yang diperoleh dari platform *Kaggle*, dengan sumber utama berasal dari hasil *web scraping* terhadap situs LinkedIn, salah satu platform terbesar untuk pencarian kerja profesional secara global. Dataset ini secara khusus menghimpun lowongan pekerjaan yang berfokus pada bidang data, seperti Data Analyst, Data Scientist, dan Data Engineer, sehingga sangat relevan untuk dianalisis menggunakan pendekatan *Natural Language Processing* (NLP) dan klasifikasi teks. Dataset ini terdiri dari 327 entri lowongan kerja, yang mencakup berbagai informasi penting terkait posisi, perusahaan, lokasi, hingga uraian deskripsi pekerjaan.

Setiap entri dalam dataset memuat informasi dalam bentuk struktur tabular dengan total 11 kolom utama. Di antaranya, kolom title berisi nama jabatan, company menunjukkan perusahaan yang membuka lowongan, location mencerminkan wilayah penempatan kerja, serta description yang berisi informasi rinci mengenai tanggung jawab, kualifikasi, dan benefit dari posisi tersebut. Kolom link dan source memberikan referensi ke halaman asli lowongan di LinkedIn. Sementara itu, date_posted mencatat tanggal unggahan setiap lowongan, yang dapat digunakan untuk melihat tren posting harian. Kolom work_type dan employment_type

disediakan namun tidak memiliki data (kosong), sehingga dikeluarkan dari proses analisis lebih lanjut.

Selain itu, peneliti menambahkan satu kolom penting bernama category yang dibuat berdasarkan nilai dalam kolom title. Kolom ini berfungsi sebagai label klasifikasi dan terbagi menjadi tiga kelas utama, yaitu Data Analyst, Data Scientist, dan Data Engineer. Pengkategorian ini dilakukan melalui proses pemetaan kata kunci dalam title, agar model klasifikasi yang dibangun dapat melakukan pelabelan otomatis terhadap deskripsi pekerjaan yang baru. Adapun rincian metadata dari dataset dapat dilihat pada Tabel 4.1 berikut:

Tabel 4. 1 Metadata Dataset LinkedIn Data Jobs

Kolom	**Deskripsi**
id	ID unik untuk setiap entri lowongan kerja
title	Judul pekerjaan, seperti *Data Analyst* atau *ML Engineer dll*
category	Kategori pekerjaan hasil klasifikasi manual: Data Analyst, Scientist, Engineer
company	Nama perusahaan yang membuka lowongan
location	Lokasi/kota penempatan pekerjaan
link	Tautan ke halaman lowongan di LinkedIn
source	Sumber data (LinkedIn)
date_posted	Tanggal posting lowongan kerja
work_type	Tipe kerja (Remote, Hybrid, Onsite) — *tidak tersedia*
employment_type	Jenis pekerjaan (Full-time, Internship, dll) — *tidak tersedia*
description	Deskripsi pekerjaan secara lengkap

4.2 Preprocessing Data

Langkah awal dalam analisis data ini adalah melakukan *data cleaning* dan preprocessing guna memastikan bahwa dataset yang digunakan berada dalam kondisi optimal untuk analisis lanjutan. Berdasarkan hasil eksplorasi awal, ditemukan bahwa dua kolom yaitu work_type dan employment_type tidak memiliki satu pun nilai yang terisi (berisi NaN seluruhnya), sehingga dihapus dari dataset. Selain itu, seluruh kolom lainnya terisi penuh tanpa adanya nilai kosong, sehingga tidak diperlukan langkah imputasi data. Proses ini menghasilkan dataset bersih dengan total 9 kolom, yang terdiri dari informasi kunci terkait identitas dan deskripsi lowongan kerja.

Langkah selanjutnya adalah melakukan normalisasi teks pada kolom description, yang merupakan sumber utama informasi pekerjaan. Proses normalisasi dilakukan melalui beberapa tahapan: (1) *case folding* untuk mengubah seluruh huruf menjadi huruf kecil, (2) penghapusan angka dan simbol menggunakan ekspresi reguler, (3) *tokenization* untuk memecah kalimat menjadi kata-kata, (4) penghapusan *stopword* bahasa Inggris menggunakan pustaka NLTK, serta (5) *stemming* menggunakan algoritma Porter untuk mereduksi kata ke bentuk dasarnya. Hasil akhir dari proses ini disimpan dalam kolom baru clean_description, yang kemudian akan digunakan untuk ekstraksi fitur dengan TF-IDF.

Sebagai tambahan, dilakukan pula proses pemetaan geografis untuk mengelompokkan nilai pada kolom location ke dalam nama negara menggunakan teknik pencocokan pola berbasis *regular expressions*. Proses ini menghasilkan

kolom country, yang memungkinkan analisis berbasis wilayah seperti tren negara dengan lowongan terbanyak. Berikut ditampilkan tiga contoh data hasil pembersihan teks pada kolom clean_description:

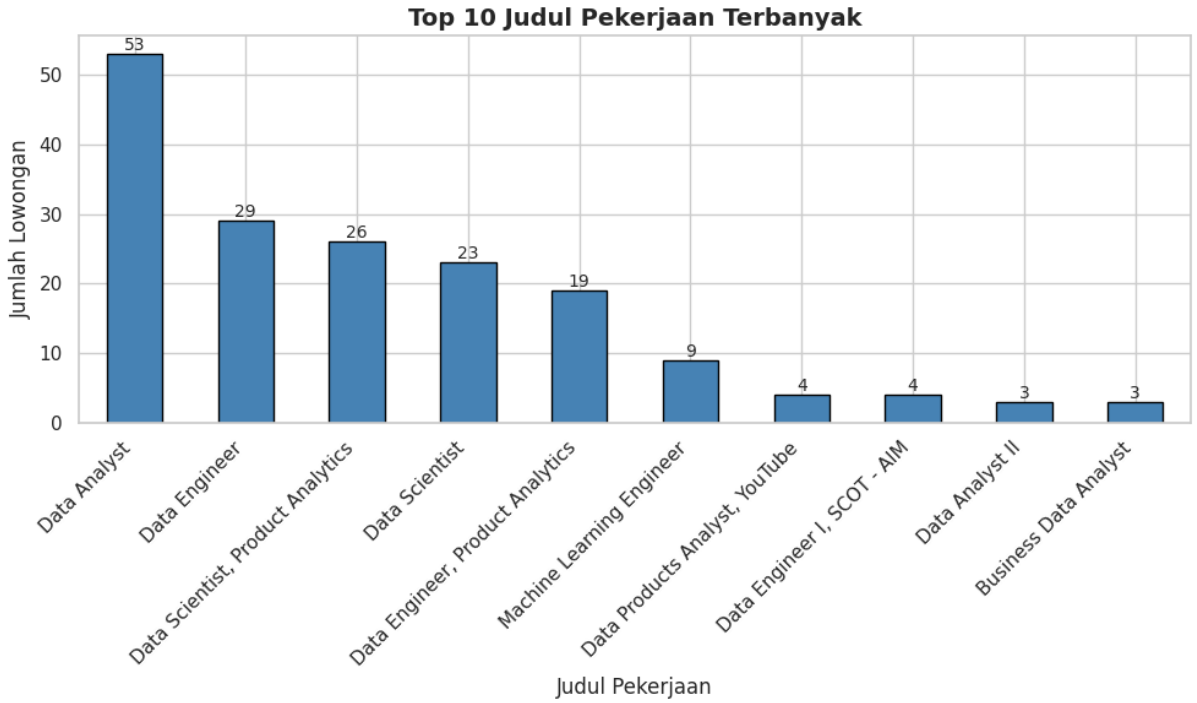
Tabel 4. 2 Contoh Deskripsi Pekerjaan Setelah *Preprocessing*

Deskripsi Asli	**Deskripsi Setelah *Preprocessing***
The Social Measurement team is a growing team ...	social measur team grow team highvis within compani respons social metr ...
About Pinterest. Millions of people around the world...	pinterest million peopl around world come platform share inspir creat ...
Snap Inc is a technology company. We believe that...	snap inc technolog compani believ camera present futur communic explor ...

4.3 Analisis Statistik Deskriptif

Analisis statistik deskriptif dilakukan untuk memperoleh pemahaman awal terhadap struktur dan distribusi data dalam dataset lowongan pekerjaan bidang data yang dianalisis. Tahapan ini penting untuk mengidentifikasi pola-pola menarik sebelum dilakukan tahapan modeling lanjutan. Visualisasi dalam bentuk grafik batang dan garis digunakan agar temuan dapat disampaikan secara visual dan mudah dipahami.

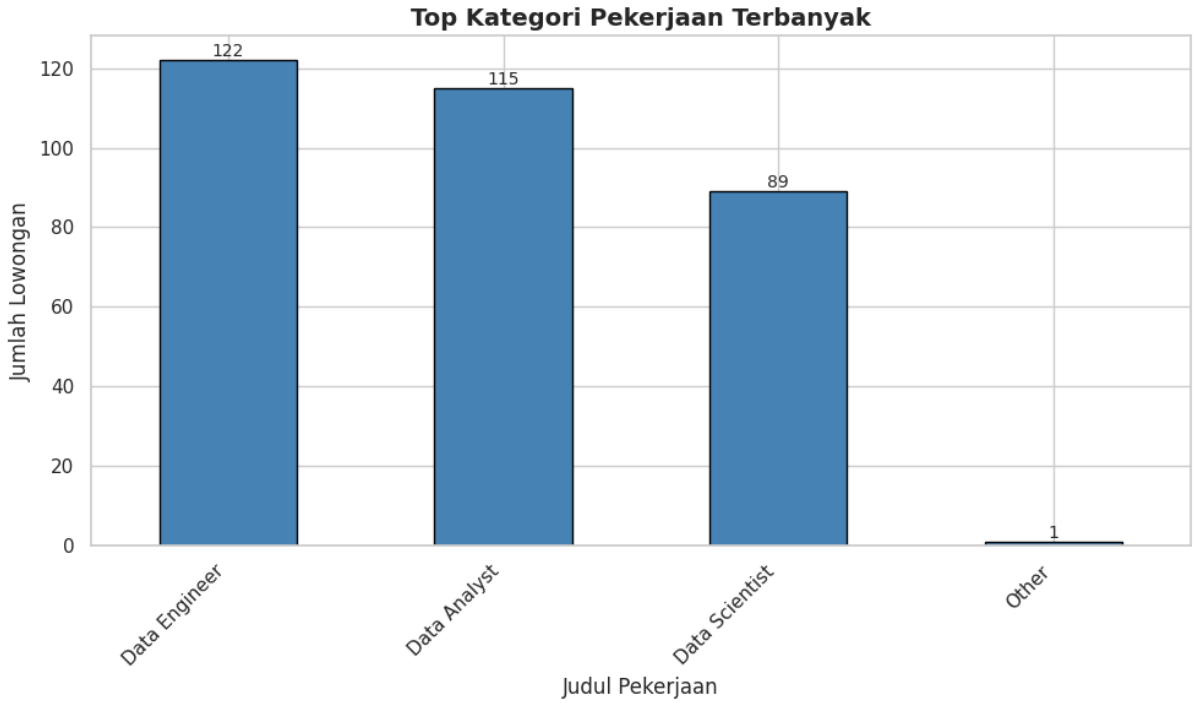
4.3.1. Top 10 Judul Pekerjaan Terbanyak



Gambar 4. 1 Bagan Top 10 Judul Pekerjaan Terbanyak

Berdasarkan hasil visualisasi, posisi Data Analyst merupakan judul pekerjaan yang paling banyak dicari oleh perusahaan, dengan total 53 lowongan. Disusul oleh Data Engineer (29 lowongan) dan Data Scientist – Product Analytics (26 lowongan). Judul-judul lainnya seperti *Machine Learning Engineer* dan *Data Products Analyst* turut muncul dalam daftar 10 besar. Hal ini menunjukkan bahwa profesi *data-driven* kini semakin terdiversifikasi, tidak hanya terbatas pada analis dan saintis, tetapi juga mencakup fungsi strategis seperti product analytics dan engineering implementation.

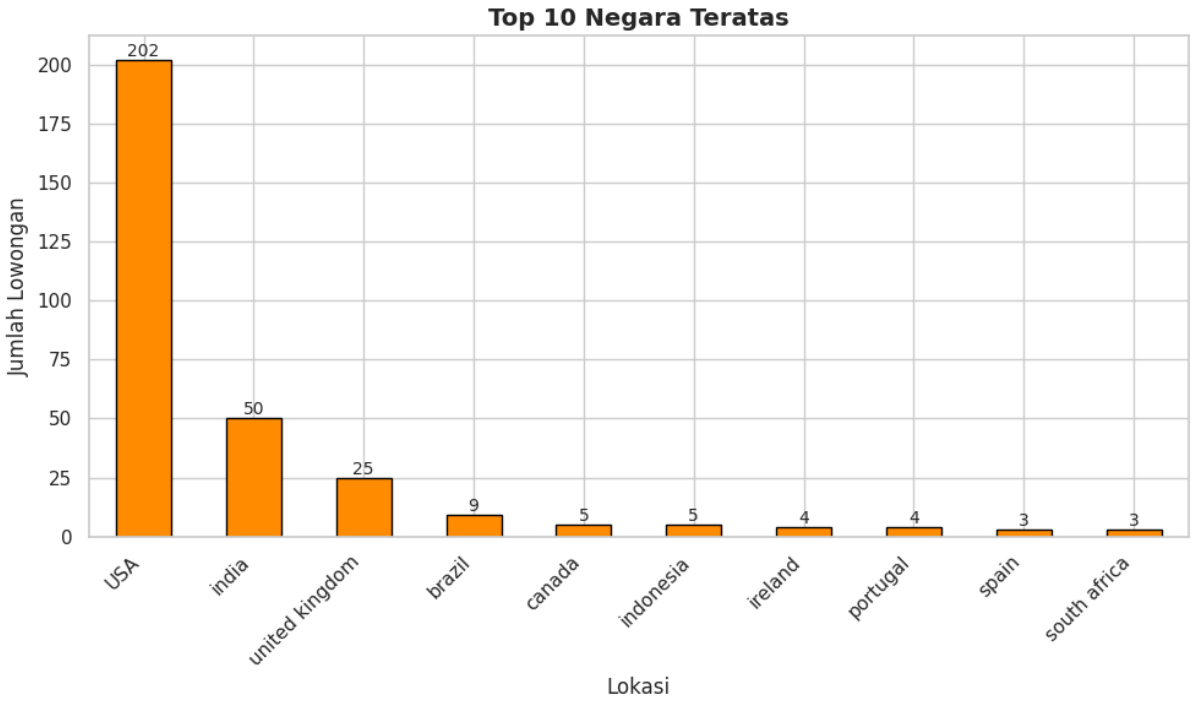
4.3.2. Top Kategori Pekerjaan Terbanyak



Gambar 4. 2 Bagan Top Kategori Pekerjaan Terbanyak

Jika dilihat dari pengelompokan kategori pekerjaan (category), Data Engineer menduduki peringkat pertama dengan 122 lowongan, disusul oleh Data Analyst (115 lowongan) dan Data Scientist (89 lowongan). Fakta ini mengindikasikan adanya kebutuhan industri yang besar terhadap talenta teknik yang tidak hanya dapat menganalisis data, tetapi juga membangun pipeline, infrastruktur, dan sistem pemrosesan data berskala besar. Ini mencerminkan tren global di mana perusahaan membutuhkan integrasi antara kemampuan analitik dan rekayasa sistem data yang solid.

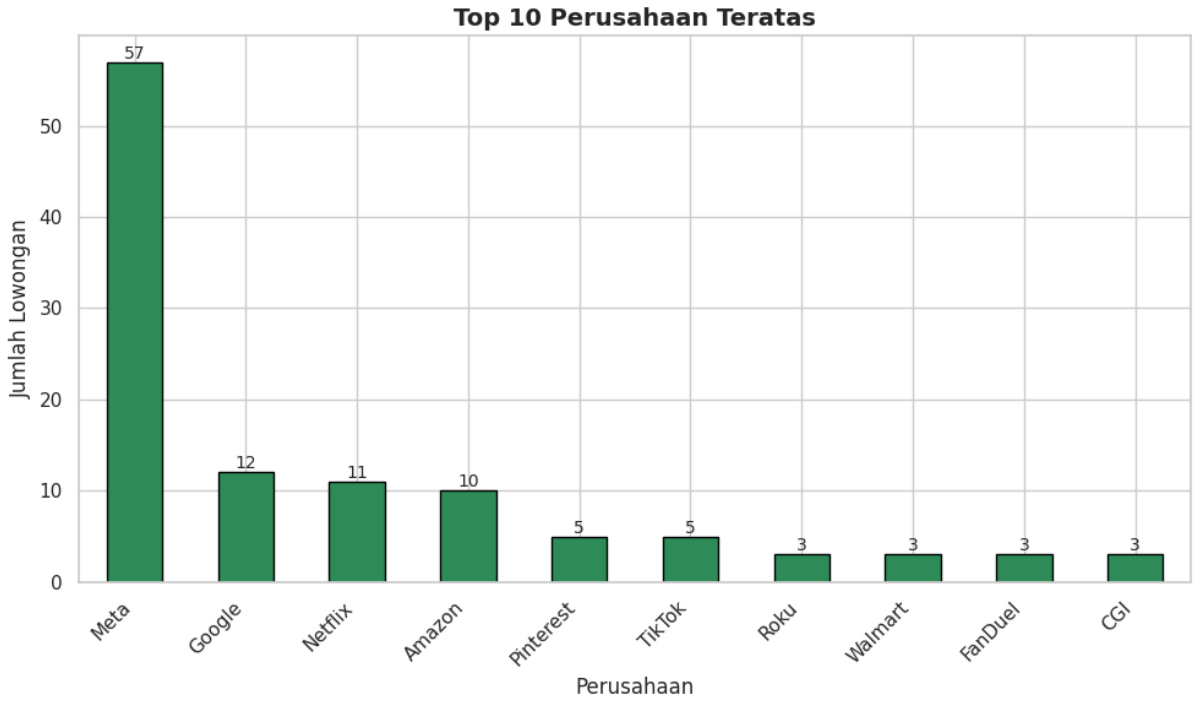
4.3.3. Top 10 Negara dengan Lowongan Terbanyak



Gambar 4. 3 Bagan Top 10 Negara dengan Lowongan Terbanyak

Dari sisi distribusi geografis, Amerika Serikat (USA) menjadi negara dengan jumlah lowongan terbanyak (202 lowongan), diikuti oleh India (50 lowongan) dan United Kingdom (25 lowongan). Hal ini dapat dipahami karena ketiga negara tersebut memiliki ekosistem teknologi yang besar dan menjadi pusat bagi perusahaan digital serta startup berbasis data. Menariknya, negara-negara seperti Brasil, Indonesia, dan Portugal juga mulai muncul sebagai kontributor, mengindikasikan bahwa adopsi teknologi data mulai menyebar secara global.

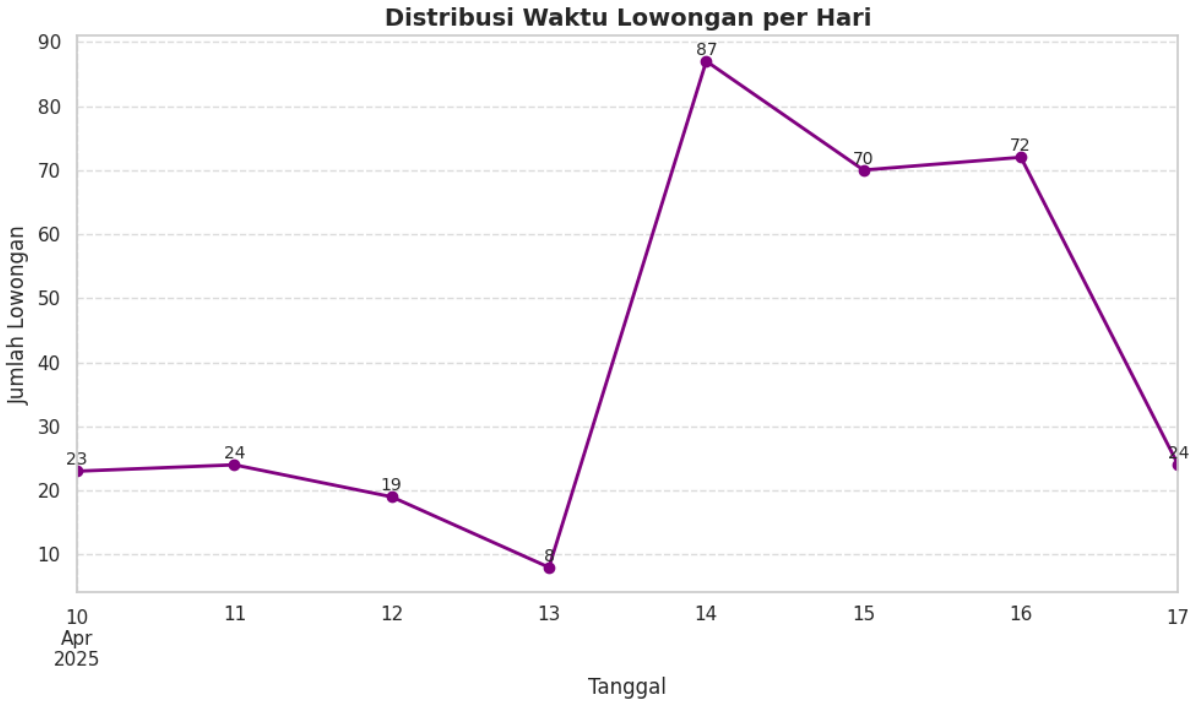
4.3.4. Top 10 Perusahaan yang Paling Aktif Merekrut



Gambar 4. 4 Bagan Top 10 Perusahaan yang Paling Aktif Merekrut

Perusahaan Meta menjadi entitas paling aktif dalam merekrut tenaga kerja bidang data, dengan total 57 lowongan, jauh mengungguli pesaingnya seperti Google (12), Netflix (11), dan Amazon (10). Ini menunjukkan betapa besarnya kebutuhan perusahaan teknologi besar terhadap talenta data untuk mengelola volume informasi pengguna yang sangat masif. Selain itu, muncul pula perusahaan seperti TikTok, Walmart, dan FanDuel, yang menandakan bahwa kebutuhan akan profesional data tidak hanya terbatas pada sektor teknologi murni, tetapi juga merambah ke industri hiburan, ritel, dan finansial.

4.3.5. Distribusi Harian Lowongan



Gambar 4. 5 Grafik Distribusi Harian Lowongan

Analisis terhadap kolom date_posted menunjukkan adanya lonjakan signifikan pada tanggal 14 April 2025 dengan 87 lowongan terposting dalam satu hari. Ini mungkin disebabkan oleh jadwal rekrutmen massal atau pembaruan data internal LinkedIn pada hari tersebut. Puncak lainnya terjadi pada tanggal 15 dan 16 April, menunjukkan periode aktif dari sisi rekrutmen. Sebaliknya, tanggal 13 April memiliki jumlah terendah (8 lowongan), yang bisa jadi berkaitan dengan hari libur atau akhir pekan. Pola ini dapat dimanfaatkan oleh pencari kerja untuk mengetahui waktu optimal dalam melamar pekerjaan secara daring.

4.4 Penerapan Teknik Text Mining

4.4.1 Transformasi TF-IDF sebagai Representasi Numerik

Untuk mengubah teks deskripsi pekerjaan menjadi bentuk numerik yang dapat diproses oleh algoritma klasifikasi, digunakan teknik Term Frequency-Inverse Document Frequency (TF-IDF). Metode ini mengukur seberapa penting suatu kata dalam dokumen tertentu relatif terhadap seluruh korpus. Kata-kata yang sering muncul dalam satu dokumen tetapi jarang ditemukan di dokumen lain akan memiliki bobot lebih tinggi. Proses transformasi menghasilkan matriks fitur berdimensi 327 baris (dokumen) dan 100 kolom (kata unik dengan skor tertinggi). Tabel 4.4.1 menampilkan cuplikan lima baris pertama dari hasil representasi TF-IDF.

Tabel 4. 3 Cuplikan Matriks TF-IDF (lima entri pertama)

abil	across	analysi	analyt	applic	base	benefit	build	busi	collabor
0.0000	0.0000	0.0715	0.1382	0.1506	0.1065	0.0974	0.1141	0.0260	0.0000
0.0000	0.0000	0.0715	0.1382	0.1506	0.1065	0.0974	0.1141	0.0260	0.0000
0.0000	0.0000	0.0715	0.1382	0.1506	0.1065	0.0974	0.1141	0.0260	0.0000
0.0000	0.0000	0.0715	0.1382	0.1506	0.1065	0.0974	0.1141	0.0260	0.0000

0.0582	0.0000	0.1071	0.0828	0.2255	0.1595	0.0486	0.0427	0.0780	0.1864

Lebih lanjut, dilakukan perhitungan rata-rata skor TF-IDF dari seluruh dokumen untuk mengidentifikasi kata-kata yang paling penting secara umum. Kata "data" memiliki skor tertinggi, diikuti oleh “experi”, “work”, dan “team” yang secara semantik menggambarkan fokus utama industri data. Informasi ini berguna untuk menyusun model klasifikasi serta menganalisis tren kata dalam lowongan pekerjaan.

Tabel 4. 4 Lima Kata dengan Skor Rata-Rata TF-IDF Tertinggi

Kata	**Rata-rata Skor TF-IDF**
data	0.3642
experi	0.1682
work	0.1189
team	0.1105
product	0.1093

4.4.2 Visualisasi Word Cloud

Gambar 4. 6 Word Cloud dari Deskripsi Pekerjaan

Visualisasi Word Cloud pada Gambar 4.4.2 memperkuat dominasi kata “data”, yang tampil paling besar di antara kata lainnya. Selain itu, kata-kata seperti “develop”, “team”, “analyt”, dan “build” muncul dengan ukuran besar, mengindikasikan bahwa perusahaan sangat menekankan keterlibatan dalam pembangunan sistem dan kerja tim. Menariknya, kata “meta” juga terlihat menonjol, kemungkinan besar karena perusahaan Meta memiliki kontribusi signifikan dalam jumlah lowongan yang tersedia dalam dataset ini.

Kehadiran kata-kata seperti “technolog”, “engin”, “model”, dan “skill” memberi gambaran bahwa sebagian besar deskripsi memuat persyaratan teknis yang cukup kompleks. Ini menandakan bahwa rekrutmen talenta data tidak hanya mempertimbangkan aspek konseptual, tetapi juga menekankan kemampuan

implementatif. Word Cloud ini memberikan representasi visual yang kuat mengenai nuansa umum dalam tuntutan kerja di bidang data science dan engineering.

4.4.3 Frekuensi Kata Paling Umum dalam Deskripsi Pekerjaan

Analisis frekuensi kata dilakukan terhadap hasil tokenisasi dan pembersihan teks, yang mencakup 327 deskripsi pekerjaan. Kata yang paling sering muncul adalah "data" dengan jumlah kemunculan sebanyak 4.576 kali, yang mencerminkan fokus utama industri ini. Diikuti oleh "experi" (1.948 kali), "work" (1.287 kali), "team" (1.224 kali), dan "product" (1.158 kali), frekuensi tinggi kata-kata ini menunjukkan bahwa deskripsi pekerjaan di bidang data sangat berorientasi pada kolaborasi tim, pengalaman kerja, dan pengembangan berbasis produk. Analisis ini memberikan landasan yang kuat untuk pemilihan fitur dalam proses klasifikasi teks selanjutnya.

Tabel 4. 5 Dua Puluh Kata dengan Frekuensi Tertinggi

Kata	**Frekuensi**
data	4,576
experi	1,948
work	1,287
team	1,224

product	1,158
busi	1,131
build	1,007
develop	911
engin	906
analyt	869
applic	769
meta	769
skill	722
includ	698
year	685
use	676
model	628

process	608
opportun	599
benefit	582

4.5 Model Klasifikasi Jenis Pekerjaan

4.5.1 Pendekatan Klasifikasi dengan Decision Tree

Pendekatan klasifikasi pada penelitian ini menggunakan algoritma Decision Tree, yang dipilih karena kemampuannya dalam menangani data kategorikal dan memberikan interpretasi model yang cukup mudah dipahami. Decision Tree bekerja dengan memetakan aturan-aturan logika dalam bentuk struktur pohon, yang secara hierarkis membagi data berdasarkan atribut paling diskriminatif. Dalam konteks ini, fitur-fitur yang diperoleh dari representasi teks melalui TF-IDF digunakan sebagai masukan untuk model. Tujuan utamanya adalah memprediksi jenis pekerjaan berdasarkan isi deskripsi lowongan kerja, yang telah melalui proses pembersihan dan transformasi teks sebelumnya.

4.5.2 Preprocessing Label Kategori

Label pekerjaan yang tersedia dalam dataset awal berasal dari atribut title, yang sangat bervariasi. Untuk menyederhanakan klasifikasi, dilakukan pengelompokan secara manual terhadap judul pekerjaan ke dalam tiga kategori utama yaitu: Data

Analyst, Data Engineer, dan Data Scientist. Label yang memiliki frekuensi sangat rendah atau tidak relevan digolongkan ke dalam kategori "Other", namun pada tahap modeling hanya label dengan jumlah kemunculan lebih dari satu yang digunakan. Langkah ini bertujuan untuk meningkatkan keseimbangan data dan keakuratan hasil pelatihan model.

4.5.3 Evaluasi Model

Evaluasi performa model Decision Tree dilakukan untuk mengukur efektivitas klasifikasi berdasarkan deskripsi pekerjaan yang telah diproses menjadi fitur numerik. Empat metrik utama digunakan dalam proses evaluasi: akurasi, precision, recall, dan F1-score. Hasil yang diperoleh menunjukkan bahwa model memiliki akurasi sebesar 80,3%, yang berarti sekitar 80 dari 100 prediksi model berhasil mengklasifikasikan jenis pekerjaan dengan benar. Ini menunjukkan tingkat ketepatan klasifikasi yang cukup tinggi untuk konteks teks deskripsi pekerjaan yang bersifat kompleks dan penuh variasi.

Jika dilihat lebih rinci, nilai precision tertinggi dicapai oleh kelas Data Analyst sebesar 0.87, menandakan bahwa 87% dari prediksi yang diklasifikasikan sebagai Data Analyst memang benar-benar termasuk ke dalam kelas tersebut. Hal ini menunjukkan bahwa model mampu menghindari banyak kesalahan tipe I (false positive) pada kelas ini. Sementara itu, kelas Data Engineer dan Data Scientist memiliki precision masing-masing 0.79 dan 0.74, yang walaupun sedikit lebih rendah, masih dalam kategori baik. Nilai recall menggambarkan seberapa besar proporsi kelas yang berhasil dikenali dengan benar oleh model. Model memberikan

recall sebesar 0.87 untuk Data Analyst, 0.76 untuk Data Engineer, dan 0.78 untuk Data Scientist, yang menunjukkan bahwa model juga cukup andal dalam menangkap data dari ketiga kategori tersebut.

Nilai F1-score, yang merupakan harmonisasi antara precision dan recall, juga menunjukkan performa seimbang: 0.87 untuk Data Analyst, 0.78 untuk Data Engineer, dan 0.76 untuk Data Scientist. Ini berarti model tidak hanya akurat, tetapi juga memiliki keseimbangan dalam mengidentifikasi dan membedakan antar kategori pekerjaan. Nilai rata-rata makro dan tertimbang dari keempat metrik ini konsisten di angka 0.80, menegaskan bahwa tidak ada kelas yang sangat dominan maupun terlalu minor dalam hasil klasifikasi.Tabel berikut merangkum nilai metrik evaluasi tersebut:

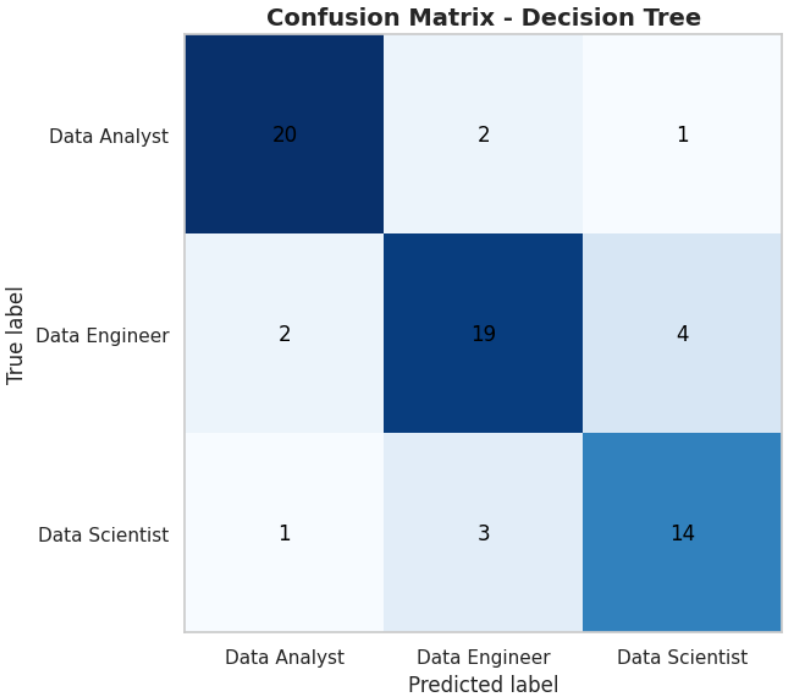
Tabel 4. 6 Hasil Evaluasi Metrik Klasifikasi (*Precision, Recall, F1-Score*)

Kelas	**Precision**	**Recall**	**F1-Score**	**Jumlah Data (Support)**
Data Analyst	0.87	0.87	0.87	23
Data Engineer	0.79	0.76	0.78	25
Data Scientist	0.74	0.78	0.76	18
Macro Avg	0.80	0.80	0.80	66
Weighted Avg	0.80	0.80	0.80	66

Akurasi	—	—	**0.803**	66

Untuk memperdalam pemahaman terhadap pola kesalahan yang terjadi selama proses klasifikasi, digunakan Confusion Matrix, yang divisualisasikan pada Gambar 4.5.2 berikut. Matriks ini menunjukkan jumlah prediksi benar (diagonal) dan prediksi salah (di luar diagonal) untuk setiap pasangan kelas. Sebagai contoh, dari total 23 lowongan kerja dengan label Data Analyst, model berhasil memprediksi 20 dengan benar, sementara 2 salah diklasifikasikan sebagai Data Engineer, dan 1 sebagai Data Scientist. Ini menunjukkan adanya ambiguitas yang masih terjadi antara deskripsi pekerjaan Data Analyst dan dua kelas lainnya, meskipun secara umum performa prediksi cukup tinggi.

Khusus pada kelas Data Engineer, terdapat 4 kasus yang salah diklasifikasikan menjadi Data Scientist. Hal ini mencerminkan adanya overlap deskriptif antara kedua profesi, seperti penggunaan kata-kata seperti “pipelin”, “insight”, atau “analysi” yang bersifat generik. Adapun untuk kelas Data Scientist, sebanyak 14 dari 18 data berhasil diprediksi dengan benar, dengan 3 kasus salah klasifikasi tersebar ke dua kelas lainnya. Hal ini menunjukkan bahwa deskripsi pekerjaan Data Scientist cenderung lebih sulit dipisahkan dari kategori lain dalam konteks teks alami.

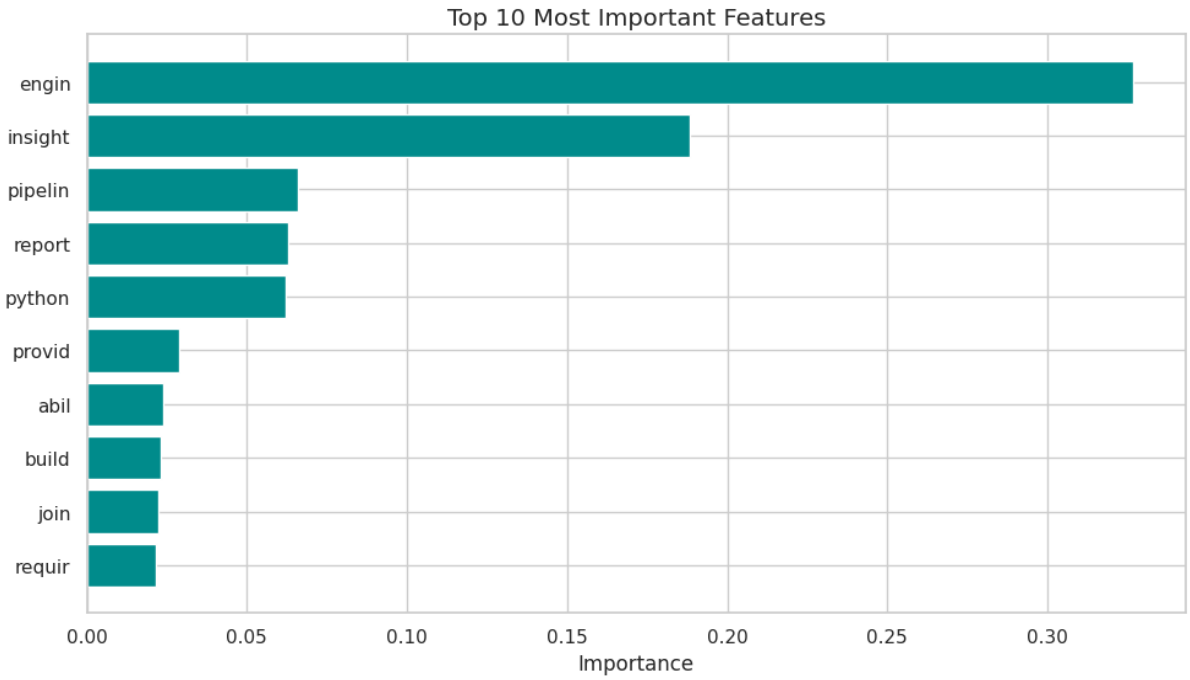


Gambar 4. 7 Confusion Matrix Model Decision Tree

Selain itu, dilakukan analisis terhadap *feature importance* untuk mengetahui kata-kata yang paling berkontribusi dalam proses klasifikasi. Hasil dari metode ini menunjukkan bahwa kata "engin" (yang berasal dari *engineer*) menjadi fitur paling dominan dalam membedakan kategori pekerjaan, diikuti oleh "insight", "pipelin", dan "python". Kata-kata ini memang sangat identik dengan pekerjaan berbasis engineering dan data science yang menekankan penggunaan teknologi dan pemrosesan data skala besar.

Visualisasi kontribusi fitur dapat dilihat pada Gambar 4.8, yang memperlihatkan 10 fitur terpenting berdasarkan skor pentingnya. Informasi ini sangat bermanfaat tidak hanya dalam interpretasi model, tetapi juga dapat digunakan oleh perusahaan atau platform rekrutmen untuk menyusun deskripsi

pekerjaan yang lebih tepat sasaran. Dengan mengutamakan kata-kata yang relevan dan spesifik, proses pencocokan antara lowongan dan kandidat dapat lebih optimal.



Gambar 4. 8 Top 10 Most Important Features dari Model Decision Tree

Grafik 4.8 menjabarkan hasil tentang fitur mana yang paling berpengaruh terhadap hasil prediksi dalam model Decision Tree. Kata-kata seperti engin, insight, dan pipelin sangat berperan dalam membentuk keputusan model, yang bisa berguna untuk interpretabilitas atau seleksi fitur.

4. 6 Hasil Analisis

Penelitian ini berhasil memanfaatkan pendekatan text mining untuk mengklasifikasikan jenis pekerjaan di bidang data berdasarkan deskripsi lowongan kerja yang diambil dari LinkedIn. Dengan total 327 entri, dataset dibersihkan dan diproses

menggunakan teknik Natural Language Processing (NLP) seperti tokenization, stopword removal, dan stemming. Tahapan ini sesuai dengan pendekatan yang digunakan dalam studi Das et al. (2023), yang menekankan pentingnya preprocessing dalam meningkatkan kualitas representasi teks untuk klasifikasi. Dalam penelitian tersebut, preprocessing menjadi fondasi utama sebelum penerapan metode TF-IDF dan klasifikasi lanjutan.

Transformasi deskripsi lowongan menjadi representasi numerik dilakukan menggunakan TF-IDF, menghasilkan 100 fitur yang merepresentasikan kata-kata paling bermakna dari segi bobot informasi. Hasil TF-IDF menyoroti kata-kata seperti data, engineer, insight, dan pipeline sebagai yang paling dominan. Temuan ini memperkuat hasil Wen Zhang et al. (2011), yang menunjukkan bahwa TF-IDF efektif dalam mengekstraksi fitur penting dari teks pendek, terutama dalam konteks klasifikasi dokumen. Namun, Zhang juga mencatat bahwa untuk data dengan semantik yang kompleks, seperti deskripsi pekerjaan yang bisa memiliki makna kontekstual tinggi, pendekatan seperti LSI (Latent Semantic Indexing) bisa memberikan akurasi yang lebih baik.

Model klasifikasi yang digunakan adalah Decision Tree Classifier, yang dipilih karena sifatnya yang interpretatif dan kemampuannya menangani fitur kategorikal maupun numerik. Model ini mencapai akurasi 80.3% dan menunjukkan performa yang cukup stabil di ketiga kelas utama: Data Analyst, Data Engineer, dan Data Scientist. Precision tertinggi diraih oleh kelas Data Analyst, sementara Data Scientist dan Data Engineer menunjukkan beberapa tumpang tindih dalam klasifikasi. Hal ini kemungkinan disebabkan oleh kemiripan terminologi teknis dalam deskripsi pekerjaan, seperti penggunaan kata develop, model, dan analysis, yang muncul di berbagai kategori. Hasil

ini sejalan dengan temuan Das et al. (2023), yang menemukan bahwa Decision Tree bisa menghasilkan performa yang kompetitif, meskipun Random Forest cenderung lebih unggul dari segi akurasi dan generalisasi.

Ketika dibandingkan dengan studi S. Sharma et al. (2020) yang menggunakan TF-IDF dan Random Forest untuk klasifikasi lowongan pekerjaan palsu (fraud job detection), akurasi yang dicapai dapat mencapai hingga 94%. Ini menunjukkan bahwa kombinasi TF-IDF dan metode ensemble learning bisa menjadi alternatif yang menjanjikan untuk skenario klasifikasi teks berbasis pekerjaan. Oleh karena itu, meskipun Decision Tree cukup efektif untuk studi ini, pengembangan model menggunakan Random Forest atau Gradient Boosting dapat diteliti lebih lanjut untuk meningkatkan performa klasifikasi, terutama dalam menangani overlap semantik antar kategori.

Dari segi fitur penting (feature importance), penelitian ini mengidentifikasi bahwa kata-kata seperti engineer, insight, dan python memiliki pengaruh besar terhadap proses klasifikasi. Ini sejalan dengan kebutuhan industri saat ini yang mengharuskan calon profesional di bidang data untuk memiliki pemahaman mendalam tentang alat dan konsep teknis. Studi oleh P. Albrecht et al. (2022) juga menunjukkan pentingnya keterampilan teknis spesifik, terutama Python dan SQL, dalam pekerjaan yang berorientasi data. Oleh karena itu, hasil feature importance tidak hanya bermanfaat untuk klasifikasi, tetapi juga dapat digunakan oleh perekrut untuk menyusun deskripsi pekerjaan yang lebih tepat sasaran dan menarik kandidat yang relevan.

Secara keseluruhan, pendekatan text mining dengan TF-IDF dan Decision Tree pada data deskripsi pekerjaan LinkedIn telah memberikan hasil yang cukup baik dan

valid, serta memiliki koherensi dengan berbagai temuan dalam literatur akademik. Namun, ada ruang untuk peningkatan baik dari sisi model maupun pendekatan representasi fitur, seperti menggunakan word embeddings (Word2Vec, BERT) untuk menangkap makna kontekstual yang lebih dalam. Studi lanjutan dapat memperkaya metodologi ini dengan membandingkan performa berbagai algoritma klasifikasi serta menganalisis faktor-faktor lain seperti perusahaan, lokasi, atau senioritas sebagai variabel prediktif tambahan.

4.7 Hubungan antara hasil penelitian dalam Islam

1. Allah memerintahkan untuk saling tolong-menolong

وَتَعَاوَنُوْا عَلَى الْبِرِّ وَالتَّقْوٰىۖ وَلَا تَعَاوَنُوْا عَلَى الْاِثْمِ وَالْعُدْوَانِ

"Dan tolong-menolonglah kamu dalam (mengerjakan) kebajikan dan takwa, dan jangan tolong-menolong dalam berbuat dosa dan permusuhan(Qura'an, Surah,AlMa'idah 02)

"Allah Ta'ala memerintahkan hamba-hamba-Nya yang beriman untuk saling tolong-menolong dalam melakukan kebaikan, yaitu kebajikan (al-birr), dan meninggalkan kemungkaran, yaitu ketakwaan. Dan Dia melarang mereka saling membantu dalam kebatilan serta bekerja sama dalam dosa dan pelanggaran terhadap syariat(tafsir ibnu kathir, 1420H,2000)

2. Manfaat penelitian ini untuk masyarakat adalah membantu memudahkan untuk mencari pekerjaan yang sesuai dengan keterampilan.

مَنْ دَلَّ عَلَى خَيْرٍ فَلَهُ مِثْلُ أَجْرِ فَاعِلِه

"Barang siapa yang menunjukkan kepada kebaikan, maka ia mendapatkan pahala seperti pahala orang yang mengerjakannya."(HR. Muslim, no. 1893).

“Hadis ini menunjukkan bahwa siapa pun yang menunjukkan atau mengajarkan kebaikan, maka ia akan mendapatkan pahala yang sama seperti orang yang melakukannya. Ini merupakan bentuk keadilan dan kemurahan Allah dalam memberi balasan kepada orang-orang yang menyebarkan petunjuk dan kebaikan.” (Syarh Muslim – Imam an-Nawawi, Juz 13).

BAB V

KESIMPULAN

5. 1 Kesimpulan

Berdasarkan hasil penelitian yang telah dilakukan, maka dapat diambil beberapa kesimpulan sebagai berikut:

1. Dataset LinkedIn yang digunakan berhasil diproses dan dianalisis menggunakan metode text mining. Dataset terdiri dari 327 lowongan kerja di bidang data (Data Analyst, Data Scientist, dan Data Engineer) dan telah dibersihkan melalui proses preprocessing teks yang komprehensif.
2. Penerapan teknik TF-IDF efektif dalam mengekstrak fitur penting dari deskripsi pekerjaan. Kata-kata seperti “data”, “experi”, “work”, “team”, dan “product” muncul sebagai kata kunci dominan dalam industri data, yang mencerminkan kebutuhan atas keterampilan kolaboratif dan teknikal.
3. Analisis statistik deskriptif menunjukkan bahwa Data Engineer merupakan kategori pekerjaan yang paling banyak dibuka, dengan konsentrasi lowongan di negara seperti Amerika Serikat dan India, serta perusahaan seperti Meta dan Google sebagai perekrut utama.
4. Model klasifikasi Decision Tree berhasil mengelompokkan jenis pekerjaan dengan tingkat akurasi mencapai 80,3%. Model menunjukkan performa yang baik, khususnya dalam mengklasifikasikan pekerjaan Data Analyst (precision dan recall mencapai 0.87), dan mengidentifikasi fitur-fitur penting seperti “engin”, “insight”, dan “pipelin”.

5. Visualisasi seperti Word Cloud dan feature importance plot mampu memberikan gambaran intuitif mengenai kebutuhan keterampilan dan kata-kata kunci yang paling relevan, yang dapat dimanfaatkan oleh pencari kerja, penyusun kurikulum pendidikan, maupun perusahaan rekrutmen

5.2 Saran

Berdasarkan temuan penelitian ini, beberapa saran dapat diberikan:

1. Dataset yang lebih besar dan lebih beragam dari berbagai sektor industri di luar bidang data akan memperkuat generalisasi hasil penelitian.
2. Perlu dilakukan pengujian lebih lanjut dengan algoritma klasifikasi lain, seperti Random Forest, SVM, atau XGBoost untuk melihat peningkatan performa model.
3. Penggabungan analisis sentimen terhadap deskripsi pekerjaan bisa menambah dimensi baru dalam pemahaman iklim kerja yang ditawarkan.
4. Pengembangan sistem berbasis web atau dashboard visualisasi dapat membantu pengguna akhir seperti mahasiswa atau profesional untuk menjelajahi tren keterampilan secara interaktif.

DAFTAR PUSTAKA

Addiga, A., & Bagui, S. (2022). Sentiment analysis on twitter data using term frequency-inverse document frequency. Journal of computer and communications, 10(8), 117-128.

Albrecht, P., Khadangi, A., & Broeck, W. V. (2022). *A skill-based analysis of job postings in data-related occupations using natural language processing*. Journal of Labor Market Research, **56**(1), 1–18. https://doi.org/10.1186/s12651-022-00314-z

Anggina, S., Setiawan, N. Y., & Bachtiar, F. A. (2022). Analisis Ulasan Pelanggan Menggunakan Multinomial Naïve Bayes Classifier dengan Lexicon-Based dan TF-IDF Pada Formaggio Coffee and Resto. @ is The Best: Accounting Information Systems and Information Technology Business Enterprise, 7(1), 76-90.

Arif, F. D. U. (2023). Perbandingan kinerja algoritma Random forest, Xgboost dan Lightgbm dalam klasifikasi emosi komentar Reddit (Bachelor's thesis, Fakultas Sains dan Teknologi UIN Syarif HIdayatullah Jakarta).

Ayu, R. (2020). *Introduction to digital marketing analytics*. LinkedIn Learning. https://www.linkedin.com/learning

Berry, M. W., & Kogan, J. (Eds.). (2010). *Text mining: Applications and theory*. Wiley.

Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python*. O'Reilly Media.

Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python*. O'Reilly Media.

Case, D. O. (2013). *Information behavior*. LinkedIn Learning. https://www.linkedin.com/learning

Chandler, M. (n.d.). *Leadership communication strategies*. LinkedIn Learning. https://www.linkedin.com/learning

Charbuty, B., & Abdulazeez, A. (2021). Classification based on decision tree algorithm for machine learning. Journal of applied science and technology trends, 2(01), 20-28.

Cheng, Y., Yu, Z., Hu, J., & Yang, M. (2022, October). A Chinese short text classification method based on TF-IDF and gradient boosting decision tree. In 2022 International Conference on Image Processing, Computer Vision and Machine Learning (ICICML) (pp. 164-168). IEEE.

Costa, V. G., & Pedreira, C. E. (2023). Recent advances in decision trees: An updated survey. Artificial Intelligence Review, 56(5), 4765-4800.

Das, S., Roy, P., & Tripathy, A. (2023). *An effective classification technique for job description using machine learning approaches*. International Journal of Data Science and Analytics, **15**(2), 87–100. https://doi.org/10.1007/s41060-022-00363-w

Feldman, R., & Sanger, J. (2007). *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. Cambridge University Press.

Fitria, A. (2024). Sistem Rekomendasi Pekerjaan Menggunakan Pendeketan Content-Based Filtering (Doctoral dissertation, Universitas Islam Negeri Maulana Malik Ibrahim Malang).

Han, J., Kamber, M., & Pei, J. (2012). *Data Mining: Concepts and Techniques* (3rd ed.). Morgan Kaufmann.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer.

Kurniawan, M. R., Pratiwi, O. N., & Hamami, F. (2025). KLASIFIKASI SOAL

Kurniawati, K. (2024). Klasifikasi data mahasiswa lampau menggunakan metode decision tree dan support vector machine (Doctoral dissertation, Universitas Islam Negeri Maulana Malik Ibrahim).

Kusnaya, W. A., Cahyana, Y., & Juwita, A. R. (2025). Penerapan Metode Naive Bayes Multinomial dan Complement dalam Membandingkan Tingkat Akurasi terhadap Analisis Sentimen Kurikulum Merdeka. Scientific Student Journal for Information, Technology and Science, 6(1), 62-69.

Larsen, R. (2020). *Data science foundations: Fundamentals*. LinkedIn Learning. https://www.linkedin.com/learning

LinkedIn Corporation. (2024). *About LinkedIn*. Diakses dari https://about.linkedin.com/

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.

Nafis, N. S. M., & Awang, S. (2021). An enhanced hybrid feature selection technique using term frequency-inverse document frequency and support vector machine-recursive feature elimination for sentiment classification. Ieee Access, 9, 52177-52192.

Nissa, M. F., Syaadah, I., Kurniasih, S., Nurlaelah, N., Asyidiq, M. Z., & Julianto, I. R. (2025). Urgensi Literasi Teknologi Informasi dan Komunikasi dalam

Pembelajaran Bahasa Indonesia di Sekolah Dasar. Jurnal Cahaya Edukasi, 3(1), 15-18.

Putra, A., & Lestari, D. (2021). Clustering Job Vacancy Based on Required Skills Using TF-IDF and K-Means. *International Conference on Data Science and Information Technology*, 55–60.

Quinlan, J. R. (1986). Induction of Decision Trees. *Machine Learning*, 1(1), 81–106.

Rahayu, I. P., Fauzi, A., & Indra, J. (2022). Analisis Sentimen Terhadap Program Kampus Merdeka Menggunakan Naive Bayes Dan Support Vector Machine. Jurnal Sistem Komputer Dan Informatika (JSON) Hal, 25-38.

Ramos, J. (2003). Using TF-IDF to Determine Word Relevance in Document Queries. In *Proceedings of the First Instructional Conference on Machine Learning*.

Ramos, J. (2003). Using TF-IDF to determine word relevance in document queries. *Proceedings of the First Instructional Conference on Machine Learning (ICML 2003)*.

Salin, D. (2017). *Managing workplace conflict*. LinkedIn Learning. https://www.linkedin.com/learning

Sari, B., Nugroho, D., & Rahman, A. (2020). Text Classification of Job Vacancies Using Naive Bayes Algorithm. *Jurnal Teknologi Informasi dan Ilmu Komputer*, 7(1), 14–20.

Sari, E. R. (2020). Text preprocessing untuk klasifikasi teks: Studi kasus pada analisis sentimen. *Jurnal Informatika dan Sistem Informasi*, *6*(2), 120–128.

Sebastiani, F. (2002). Machine Learning in Automated Text Categorization. *ACM Computing Surveys*, 34(1), 1–47.

Sharma, S., Pal, M., & Singh, R. (2020). *Fake job detection using machine learning approach*. Procedia Computer Science, **173**, 370–378. https://doi.org/10.1016/j.procs.2020.06.043

Singgalen, Y. A. (2023). Penerapan Metode CRISP-DM dalam Klasifikasi Data Ulasan Pengunjung Destinasi Danau Toba Menggunakan Algoritma Naïve Bayes Classifier (NBC) dan Decision Tree (DT). J. Media Inform. Budidarma, 7(3), 1551-1562.

Singh, J., & Tripathi, P. (2021, June). Sentiment analysis of Twitter data by making use of SVM, Random Forest and Decision Tree algorithm. In 2021 10th IEEE international conference on communication systems and network technologies (CSNT) (pp. 193-198). IEEE.

Solehuddin, M., Syafei, W. A., & Gernowo, R. (2022). Metode decision tree untuk meningkatkan kualitas rencana pelaksanaan pembelajaran dengan algoritma C4. 5. Jurnal Penelitian Dan Pengembangan Pendidikan, 6(3), 510-519.

Tafseer, N., & Ali, M. (2018). A comparative study of text classification algorithms using feature selection and TF-IDF. *International Journal of Computer Applications, 180*(42), 6–10

Wijaya, C. A., & Suryani, N. (2022). Analisis Deskripsi Pekerjaan di LinkedIn Menggunakan TF-IDF dan Decision Tree. *Jurnal Sistem Informasi*, 18(3), 211–220.

Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data Mining: Practical Machine Learning Tools and Techniques* (3rd ed.). Morgan Kaufmann.

Zhang, W., Yoshida, T., & Tang, X. (2011). *A comparative study of TF-IDF, LSI and multi-words for text classification*. Expert Systems with Applications, **38**(3), 2758–2765. https://doi.org/10.1016/j.eswa.2010.08.066

Ibn Kathir. (1420 H / 2000 M). Tafsir al-Qur'an al-'Azim. Beirut: Dar Ibn Hazm.

LAMPIRAN

id	title	category	company	location	link	source
1	Data Analyst	Data Analyst	Meta	New York, NY	https://www.linkedin.com/jobs/view/data-analyst-at-meta-4186238974	LinkedIn
2	Data Analyst	Data Analyst	Meta	San Francisco, CA	https://www.linkedin.com/jobs/view/data-analyst-at-meta-4186241553	LinkedIn
3	Data Analyst	Data Analyst	Meta	Los Angeles, CA	https://www.linkedin.com/jobs/view/data-analyst-at-meta-4186236994	LinkedIn
4	Data Analyst	Data Analyst	Meta	Washington, DC	https://www.linkedin.com/jobs/view/data-analyst-at-meta-4186237989	LinkedIn
5	Data Analyst II	Data Analyst	Pinterest	Chicago, IL	https://www.linkedin.com/jobs/view/data-analyst-ii-at-pinterest-4193349988	LinkedIn
6	Data Analyst	Data Analyst	FanDuel	New York, NY	https://www.linkedin.com/jobs/view/data-analyst-at-fanduel-4206047782	LinkedIn
7	Data Analyst, Production Finance Operations & Innovation	Data Analyst	Netflix	Los Angeles, CA	https://www.linkedin.com/jobs/view/data-analyst-production-finance-operations-innovation-at-netflix-4205626465	LinkedIn
8	Data Analyst - Marketing	Data Analyst	FanDuel	New York, NY	https://www.linkedin.com/jobs/view/data-analyst-marketing-at-fanduel-4138322262	LinkedIn
9	Data Analyst	Data Analyst	SBH Fashion	New York, NY	https://www.linkedin.com/jobs/view/data-analyst-at-sbh-fashion-4168179268	LinkedIn
14	Data Analyst II	Data Analyst	Pinterest	New York, NY	https://www.linkedin.com/jobs/view/data-analyst-ii-at-pinterest-4193356222	LinkedIn

15	Data Analyst II	Data Analyst	Pinterest	San Francisco, CA	https://www.linkedin.com/jobs/view/data-analyst-ii-at-pinterest-4193351793	LinkedIn
16	Data Analyst, Global Partnerships & Content	Data Analyst	Meta	New York, NY	https://www.linkedin.com/jobs/view/data-analyst-global-partnerships-content-at-meta-4187153427	LinkedIn
29	Data Analyst	Data Analyst	FanDuel	Atlanta, GA	https://www.linkedin.com/jobs/view/data-analyst-at-fanduel-4206048749	LinkedIn
30	Data Analyst	Data Analyst	FinThrive	United States	https://www.linkedin.com/jobs/view/data-analyst-at-finthrive-4208994892	LinkedIn
31	Senior Data Analyst	Data Analyst	alice + olivia	New York, NY	https://www.linkedin.com/jobs/view/senior-data-analyst-at-alice-%2B-olivia-4190163049	LinkedIn
32	Data Products Analyst, YouTube	Data Analyst	Google	New York, NY	https://www.linkedin.com/jobs/view/data-products-analyst-youtube-at-google-4209686717	LinkedIn
34	Customer Relationship Management Analyst	Data Analyst	Bvlgari	New York City Metropolitan Area	https://www.linkedin.com/jobs/view/customer-relationship-management-analyst-at-bvlgari-4184835114	LinkedIn
35	Data Analyst - SQL, ERP	Data Analyst	CyberCoders	Yakima, WA	https://www.linkedin.com/jobs/view/data-analyst-sql-erp-at-cybercoders-4172595144	LinkedIn
54	Marketing Data Analyst	Data Analyst	TechHuman	United States	https://www.linkedin.com/jobs/view/marketing-data-analyst-at-techhuman-4206084851	LinkedIn
55	Data Analyst	Data Analyst	Hamilton Porter	New York, NY	https://www.linkedin.com/jobs/view/data-analyst-at-hamilton-porter-4209790671	LinkedIn
56	Analytics Associate	Data Analyst	Brooklinen	Brooklyn, NY	https://www.linkedin.com/jobs/view/analytics-associate-at-brooklinen-4189094109	LinkedIn
57	Data Analyst I	Data Analyst	Equip	United States	https://www.linkedin.com/jobs/view/data-analyst-i-at-equip-4172874483	LinkedIn

58	Data Analyst	Data Analyst	Toyota North America	Plano, TX	https://www.linkedin.com/jobs/view/data-analyst-at-toyota-north-america-4210659515	LinkedIn
59	Data & Analytics, Analyst	Data Analyst	TikTok	Los Angeles, CA	https://www.linkedin.com/jobs/view/data-analytics-analyst-at-tiktok-4210110419	LinkedIn
60	Junior Data Analyst - Remote	Data Analyst	Lensa	United States	https://www.linkedin.com/jobs/view/junior-data-analyst-remote-at-lensa-4210127724	LinkedIn
61	Data Products Analyst, YouTube	Data Analyst	Google	Los Angeles, CA	https://www.linkedin.com/jobs/view/data-products-analyst-youtube-at-google-4209686721	LinkedIn
62	Data Products Analyst, YouTube	Data Analyst	Google	Mountain View, CA	https://www.linkedin.com/jobs/view/data-products-analyst-youtube-at-google-4209686719	LinkedIn
63	Data Products Analyst, YouTube	Data Analyst	Google	San Bruno, CA	https://www.linkedin.com/jobs/view/data-products-analyst-youtube-at-google-4209686718	LinkedIn
64	Data Analyst	Data Analyst	Sony	Culver City, CA	https://www.linkedin.com/jobs/view/data-analyst-at-sony-4208337899	LinkedIn
65	People Data Analyst	Data Analyst	CAVA	Washington DC-Baltimore Area	https://www.linkedin.com/jobs/view/people-data-analyst-at-cava-4211477818	LinkedIn
66	Customer Insights Analyst	Data Analyst	Nextdoor	New York, NY	https://www.linkedin.com/jobs/view/customer-insights-analyst-at-nextdoor-4194555805	LinkedIn
67	Data Analyst	Data Analyst	Salt	United States	https://www.linkedin.com/jobs/view/data-analyst-at-salt-4207613081	LinkedIn
68	Data Analyst	Data Analyst	InterEx Group	United States	https://www.linkedin.com/jobs/view/data-analyst-at-interex-group-4207982241	LinkedIn
69	Data Analyst I	Data Analyst	Colibri Group	United States	https://www.linkedin.com/jobs/view/data-analyst-i-at-colibri-group-4207003840	LinkedIn
70	Data Analyst Intern (Fall start)	Data Analyst	AARP	Washington, DC	https://www.linkedin.com/jobs/view/data-analyst-intern-fall-start-at-aarp-4179907608	LinkedIn