Accredited SINTA 5
72/E/KPT/2024

Jurnal Nasional Komputasi dan Teknologi Informasi (JNKTI)
Vol. xx No. xx, mm yyyy
P-ISSN 2620-8342
E-ISSN 2621-3052

KLASIFIKASI *HARD SKILL* DAN *SOFT SKILL* IT DARI IKLAN KERJA MENGGUNAKAN METODE BERBASIS *RULE-BASED* DAN *SIMILARITY MATCHING*

Hapsa Husain[1*], Wahyudin Hasyim[2], Alter Lasarudin[3], Tri Pratiwi Handayani[4]
[1,2,3,4]Prodi Sistem Informasi, Universitas Muhammadiyah Gorontalo
Jl. Prof. Dr. H. Mansoer Pateda, Desa Pentadio Timur, Kec. Telaga Biru, Kab. Gorontalo, Gorontalo 96181. Indonesia
E-mail: hapsahusain1@gmail.com

Abstrak - Hapsa Husain. 2026. Klasifikasi *Hard Skill* dan *Soft Skill* IT dari Iklan Kerja Menggunakan Metode Berbasis *Rule-Based* dan *Similarity Matching*. Skripsi, Program Studi Sistem Informasi, Fakultas Sains dan Ilmu Komputer, Universitas Muhammadiyah Gorontalo. Dibimbing oleh. Bapak Wahyudin Hasyim, M.Kom selaku pembimbing I dan Bapak Alter Lasarudin, M.Kom selaku pembimbing II.

Kesenjangan kompetensi antara lulusan perguruan tinggi dengan kebutuhan aktual industri Teknologi Informasi (TI) masih menjadi kendala utama dalam penyerapan tenaga kerja. Penelitian ini bertujuan untuk mengklasifikasikan kompetensi *hard skill* dan *soft skill* yang diekstraksi dari data iklan lowongan kerja guna memetakan kebutuhan industri secara presisi. Metodologi yang diterapkan menggunakan pendekatann yang menggabungkan metode *rule-based* untuk ekstraksi entitas kompetensi serta *similarity matching* dengan algoritma *Cosine Similarity* untuk proses pemetaan terhadap Standar Kompetensi Kerja Nasional Indonesia (SKKNI). Penelitian ini mengolah data sebanyak 701 iklan lowongan kerja IT dengan merujuk pada 2.126 unit kompetensi SKKNI yang mencakup 57 okupasi. Hasil penelitian menunjukkan bahwa penggunaan metode *Rule-Based* dan *Similarity Matching* mampu mengidentifikasi variasi kompetensi pada teks iklan kerja yang tidak terstruktur dengan performa sangat baik, dibuktikan dengan nilai akurasi sebesar 88,02%, presisi 88,09%, recall 96,92%, dan F1-score 92,29%. Berdasarkan Hasil penelitian mengungkap dominasi kebutuhan hard skill (62.0%) dengan prioritas pada bidang keamanan siber (security) dan basis data, sementara soft skill (38,0%) didominasi oleh kemampuan komunikasi dan pemecahan masalah. Kerangka kerja ini diharapkan dapat membantu institusi pendidikan dalam sinkronisasi kurikulum serta menjadi panduan bagi pencari kerja dalam memetakan kompetensi spesifik yang dibutuhkan industri.
Kata Kunci*: Hard Skill, Soft Skill,* IT, *Rule-Based, Cosine Similarity,* SKKNI, Iklan Kerja.

Abstract *- Hapsa Husain. 2026. Classification of IT Hard and Soft Skills from Job Advertisements Using Rule-Based and Similarity Matching Methods. Thesis, Information Systems Study Program, Faculty of Science and Computer Science, Muhammadiyah University of Gorontalo. Supervised by Mr. Wahyudin Hasyim, M.Kom., as my first supervisor, and Mr. Alter Lasarudin, M.Kom., as my second supervisor.*

The competency gap between university graduates and the actual needs of the Information Technology (IT) industry remains a major obstacle to workforce absorption. This study aims to classify hard and soft skill competencies extracted from job advertisement data to precisely map industry needs. The applied methodology uses a rule-based approach for extracting competency entities and similarity matching with the Cosine Similarity algorithm for mapping against the Indonesian National Work Competency Standards (SKKNI). This study processed data from 701 IT job advertisements, referring to 2,126 SKKNI competency units covering 57 occupations. The results of the study indicate that the Rule-Based and Similarity Matching method is capable of identifying competency variations in unstructured job advertisement text with excellent performance, as evidenced by an accuracy of 88.02%, a precision of 88.09%, a recall of 96.92%, and an F1-score of 92.29%. The results revealed a dominant need for hard skills (57.8%), with a priority on cybersecurity and databases, while soft skills (42.2%) were dominated by communication and problem-solving abilities. This framework is expected to assist educational institutions in curriculum synchronization and serve as a guide for job seekers in mapping specific competencies needed by the industry.
Keywords: Hard Skills, Soft Skills, IT, Rule-Based, Cosine Similarity, SKKNI, Job Advertisements.

Copyright © Jurnal Nasional Komputasi dan Teknologi Informasi (JNKTI)

Jurnal Nasional Komputasi dan Teknologi Informasi (JNKTI)
Vol. xx No. xx, mm yyyy
P-ISSN 2620-8342
E-ISSN 2621-3052

1. Pendahuluan

Pada era persaingan pasar bebas ASEAN (MEA), perguruan tinggi, khususnya pendidikan vokasi yang berorientasi pada penerapan keahlian tertentu, dituntut untuk mampu menghasilkan lulusan yang memiliki daya saing tinggi dengan lulusan dari berbagai negara. Kondisi persaingan global yang semakin ketat ini mendorong para profesional untuk terus berinovasi, kreatif, serta meningkatkan tingkat keterampilan, kompetensi, dan profesionalitas yang dimilikinya secara berkelanjutan. Dalam dunia kerja modern, penguasaan hard skill dan soft skill memiliki pengaruh yang sangat signifikan terhadap kinerja dan prestasi kerja karyawan. Kedua jenis keterampilan tersebut sama-sama memiliki peran penting dan saling melengkapi satu sama lain dalam menunjang keberhasilan individu di lingkungan kerja [1].

Klasifikasi yang digunakan pada portal lowongan kerja umumnya masih terbatas pada pengelompokan berdasarkan bidang pekerjaan, tanpa adanya penjabaran rinci mengenai keterampilan yang dibutuhkan. Oleh karena itu, untuk memperoleh pemetaan kebutuhan talenta digital yang lebih akurat, data lowongan kerja dari portal tersebut perlu diklasifikasikan terlebih dahulu guna mengidentifikasi lowongan yang secara spesifik memerlukan talenta digital. Inkonsistensi gaya penulisan serta format deskripsi pekerjaan yang bervariasi antar perusahaan menimbulkan hambatan dalam pemetaan kebutuhan industri secara cepat, baik bagi pencari kerja maupun institusi pendidikan [2].

Meskipun volume data lowongan pekerjaan tersedia secara masif (*big data*), informasi kualifikasi di dalamnya umumnya bersifat tidak terstruktur (*unstructured text*). Data iklan lowongan kerja dalam penelitian ini dihimpun melalui portal Jobstreet, yang dipilih karena reputasinya sebagai salah satu platform rekrutmen terbesar dengan volume iklan talenta digital yang signifikan di Indonesia. Penggunaan Jobstreet memungkinkan perolehan data teks kualifikasi yang representatif, mencakup berbagai level jabatan dari berbagai industri, sehingga menyediakan basis data yang kaya untuk diolah dan dipetakan ke dalam unit kompetensi SKKNI bidang Informasi dan Komunikasi. Proses ekstraksi data yang dilakukan melalui pendekatan konvensional terhadap ribuan iklan kerja tidak hanya memerlukan durasi waktu yang signifikan, tetapi juga memiliki tingkat kerentanan yang tinggi terhadap faktor kesalahan manusia (*human error*) serta bias subjektivitas.

Kompleksitas permasalahan meningkat seiring adanya inkonsistensi terminologi dan penggunaan bahasa yang beragam (*bilingual*). Satu entitas kompetensi yang sama sering kali direpresentasikan melalui variasi diksi yang berbeda. Adanya ambiguitas terminologi dalam kualifikasi pekerjaan ini memerlukan standar kompetensi nasional sebagai basis pelabelan yang objektif [2]. Oleh karena itu, penelitian ini menggunakan Standar Kompetensi Kerja Nasional Indonesia (SKKNI) Bidang Informasi dan Komunikasi sebagai acuan baku. Penggunaan standar ini krusial untuk menjamin validitas pemisahan antara kualifikasi teknis (*hard skill*) dan kemampuan perilaku (*soft skill*) agar selaras dengan parameter industri yang berlaku di Indonesia.

Sebagai solusi atas permasalahan tersebut, diperlukan sebuah kerangka kerja yang mampu mengombinasikan kepastian aturan dengan fleksibilitas komputasi bahasa. Penelitian ini mengusulkan implementasi metode berbasis *Rule-Based* untuk menjamin ketepatan pemetaan pada istilah teknis yang sudah baku dalam SKKNI, yang kemudian diintegrasikan dengan teknik *Similarity Matching*. Pemilihan metode *Similarity Matching* dengan menggunakan algoritma *Cosine Similarity* terbukti efektif dalam memberikan penilaian objektif dengan menghitung derajat kemiripan antara istilah hasil ekstraksi dengan dataset referensi yang bersumber dari Unit-Unit Kompetensi pada SKKNI. Hal ini memungkinkan sistem untuk memetakan istilah yang secara tekstual berbeda ke dalam kelas kompetensi yang sama meskipun istilah tersebut tidak tercantum secara harfiah dalam dokumen SKKNI.

Meskipun terdapat beberapa studi yang telah mengeksplorasi klasifikasi teks melalui pendekatan *Machine Learning* konvensional, riset yang menggabungkan keunggulan pengetahuan pakar pada sistem berbasis aturan (*rule-based*) dengan fleksibilitas pencocokan kemiripan (*similarity matching*) untuk identifikasi talenta digital masih sangat terbatas. Studi ini diproyeksikan mampu memberikan sumbangsih ilmiah di bidang *Text Mining* sekaligus menghadirkan solusi teknologi dalam mengotomatisasi pemetaan kualifikasi pelamar guna mewujudkan proses seleksi yang lebih transparan, objektif, dan efisien [3].

Keunggulan dari penelitian ini terletak pada penggunaan standar SKKNI sebagai parameter klasifikasi yang objektif. Berbeda dengan pendekatan pada studi terdahulu yang cenderung melakukan pelabelan kualifikasi secara umum atau subjektif, penelitian ini menawarkan diferensiasi yang lebih presisi melalui pemisahan antara *hard skill* yang merujuk pada Unit Kompetensi Inti, serta *soft skill* yang didasarkan pada Unit Kompetensi Umum dan Sikap Kerja. Dengan demikian, hasil pemetaan yang dihasilkan memiliki validitas tinggi dan relevansi yang diakui oleh standar industri nasional.

Copyright © Jurnal Nasional Komputasi dan Teknologi Informasi (JNKTI)

Jurnal Nasional Komputasi dan Teknologi Informasi (JNKTI)
Vol. xx No. xx, mm yyyy
P-ISSN 2620-8342
E-ISSN 2621-3052

2.Tinjauan Pustaka

2.1 Kualifikasi Kompetensi IT

Kompetensi atau keterampilan (*skill*) merupakan instrumen fundamental yang menopang kinerja profesional seorang individu. Kualitas keterampilan yang dimiliki berbanding lurus dengan produktivitas yang dihasilkan, sehingga perusahaan besar saat ini menempatkan pengembangan keterampilan sebagai prioritas investasi utama dalam pertumbuhan industri [4]. Dalam lanskap industri teknologi, kualifikasi ini secara spesifik diklasifikasikan ke dalam dua dimensi utama:

a. *Hard Skill*

Hard skill merupakan kemampuan yang dapat diperoleh melalui pendidikan di sekolah atau Universitas, dengan tujuan meningkatkan kapasitas intelektual yang terkait dengan mata pelajaran yang dipelajari. *Hard skill* mencerminkan perilaku dan keterampilan yang bersifat nyata atau dapat diamati secara langsung (*eksplisit*). Keterampilan ini menghasilkan output yang terlihat dan bersifat langsung (*visible* dan *immediate*).Kemampuan *hard skill* mencakup semua aspek yang berkaitan dengan penguasaan teori sebagai dasar analisis atau pengambilan keputusan. Penilaian terhadap hard skill dapat dilakukan melalui tes teknis (*technical test*) atau tes praktik (*practical test*) [5].

b. *Soft Skill*

Soft skill merupakan elemen penting yang harus dimiliki oleh mahasiswa agar mampu bersaing dan meraih keberhasilan di dunia kerja pada era digital. Keterampilan nonteknis seperti kemampuan berkomunikasi secara efektif, bekerja sama dalam tim, serta berpikir kritis menjadi faktor utama yang menentukan kesiapan dan kesuksesan mahasiswa dalam menghadapi lingkungan kerja yang semakin dinamis dan kompleks. Dalam konteks ini, media digital berperan penting sebagai sarana pengembangan *soft skill*, karena memberikan ruang bagi mahasiswa untuk berlatih, berinteraksi, serta beradaptasi dengan teknologi modern yang selaras dengan kebutuhan dan tuntutan dunia industri [6]

2.2 Landasan Standar Kompetensi Nasional

Dalam upaya menghasilkan klasifikasi kebutuhan industri yang akurat dan objektif, penelitian ini tidak hanya bertumpu pada proses ekstraksi kata kunci secara acak, tetapi mengacu pada standar resmi yang telah ditetapkan secara nasional. Penggunaan standar tersebut penting untuk mengurangi subjektivitas peneliti dalam menentukan pengelompokan suatu keterampilan ke dalam kategori tertentu. Dalam penelitian ini, dua instrumen utama yang dijadikan sebagai basis pengetahuan (knowledge base) adalah Peta Okupasi Nasional dan Standar Kompetensi Kerja Nasional Indonesia (SKKNI).

2.2.1 Peta Okupasi Nasional Bidang Teknologi Informasi dan Komunikasi (TIK)

Peta Okupasi Nasional adalah dokumen strategis yang memuat pemetaan berbagai jabatan atau peran kerja (*job roles*) pada sektor Teknologi Informasi dan Komunikasi (TIK) ke dalam jenjang kualifikasi yang terstandar. Dalam penelitian ini, Peta Okupasi digunakan sebagai kerangka klasifikasi (taksonomi) untuk mengelompokkan data iklan lowongan kerja ke dalam posisi yang lebih spesifik. Pemanfaatan Peta Okupasi bertujuan untuk mengurangi perbedaan dan ketidakkonsistenan penggunaan istilah jabatan pada portal lowongan kerja yang sering kali bervariasi antarperusahaan. Dengan mengacu pada dokumen tersebut, sistem memiliki rujukan yang jelas mengenai profil jabatan yang telah diakui secara nasional oleh industri, sehingga mempermudah proses penilaian dan validasi terhadap kesesuaian persyaratan dalam iklan lowongan kerja dengan standar profesi yang berlaku [7].

2.2.2 Standar Kompetensi Kerja Nasional Indonesia (SKKNI) Bidang Informasi Dan Komunikasi

Standar Kompetensi Kerja Nasional Indonesia (SKKNI) didefinisikan sebagai artikulasi kemampuan kerja yang mengintegrasikan aspek pengetahuan, keterampilan, serta sikap kerja yang relevan dengan tuntutan jabatan dan ketentuan perundang-undangan yang berlaku[8]. Dalam penelitian ini, SKKNI digunakan sebagai acuan atau parameter standar untuk memvalidasi istilah-istilah kompetensi yang ditemukan pada teks iklan lowongan kerja, sehingga proses klasifikasi memiliki dasar hukum dan relevansi industri yang jelas di Indonesia.

2.2.3 Mekanisme Sinkronisasi Peta Okupasi dan SKKNI

Dalam rangka membangun sistem klasifikasi yang akurat, penelitian ini menerapkan mekanisme sinkronisasi fungsional untuk mengintegrasikan profil jabatan makro pada Peta Okupasi dengan unit kompetensi mikro yang tertuang dalam SKKNI. Langkah sinkronisasi ini menjadi krusial karena terminologi kualifikasi pada iklan lowongan kerja bersifat dinamis dan sering kali merepresentasikan kebutuhan keahlian lintas bidang, sehingga tidak dapat hanya merujuk pada satu dokumen standar secara kaku.

Tahapan sinkronisasi fungsional dalam penelitian ini dilakukan melalui prosedur sebagai berikut:

Copyright © Jurnal Nasional Komputasi dan Teknologi Informasi (JNKTI)

Jurnal Nasional Komputasi dan Teknologi Informasi (JNKTI)
Vol. xx No. xx, mm yyyy
P-ISSN 2620-8342
E-ISSN 2621-3052

a. **Agregasi Kompetensi Lintas Dokumen SKKNI:** Mengingat satu peran pekerjaan (contohnya Software Developer) dalam praktiknya memerlukan kompetensi yang tersebar di beberapa dokumen standar, sistem melakukan penarikan unit kompetensi dari berbagai SKKNI yang relevan, seperti bidang Pemrograman, Basis Data, hingga Keamanan Informasi. Proses ini bertujuan untuk membentuk profil kualifikasi jabatan yang komprehensif dan sesuai dengan realitas industri.
b. **Ekstraksi Knowledge, Skill, and Attitude (KSA):** Unit-unit kompetensi diurai (decomposition) hingga ke level elemen kompetensi dan kriteria unjuk kerja untuk mengekstraksi kata kunci teknis (technical keywords). Kata kunci tersebut mencakup tiga dimensi utama kompetensi, yaitu pengetahuan (knowledge), keterampilan (skills), dan sikap kerja (attitude), yang selanjutnya digunakan sebagai basis data referensi (knowledge base) dalam logika Rule-Based.
c. **Normalisasi Terminologi Industri melalui Similarity Matching:** Untuk mengatasi kesenjangan diksi antara istilah praktis pada portal Jobstreet dengan istilah formal dalam standar nasional, dilakukan proses normalisasi dengan SKKNI sebagai jangkar referensi (anchor). Derajat kemiripan antara istilah pada iklan kerja dengan korpus kata kunci yang telah disinkronkan diukur menggunakan algoritma Cosine Similarity guna menjamin ketepatan identifikasi.
d. **Kategorisasi Kualifikasi Kemampuan:** data hasil sinkronisasi kemudian dikelompokkan ke dalam dua ranah utama:
    1. Unit Kompetensi Inti yang diidentifikasi sebagai indikator Hard Skill.
    2. Unit Kompetensi Umum dan Sikap Kerja yang diklasifikasikan sebagai indikator Soft Skill.

Melalui pendekatan sinkronisasi ini, matriks pemetaan yang dihasilkan diharapkan mampu memberikan hasil klasifikasi yang objektif, memiliki validitas ilmiah yang kuat, serta selaras dengan standar profesionalisme TIK yang diakui oleh Kementerian Ketenagakerjaan dan Kementerian Kominfo RI.

2.2.4 Matriks Pemetaan Okupasi dan Unit Kompetensi

Penerapan Peta Okupasi dan SKKNI dalam sistem klasifikasi diilustrasikan melalui hubungan antara peran pekerjaan (*job roles*) dengan unit kompetensi teknis yang spesifik. Tabel.1 menunjukkan contoh bagaimana deskripsi iklan lowongan kerja dipetakan ke dalam korpus data sistem sebagai acuan klasifikasi.

Table 1.Contoh Pemetaan Okupasi, SKKNI, dan Kata Kunci Sistem

Jabatan	**Kode Unit SKKNI**	**Judul Unit Kompetensi**	**Representasi Kata Kunci (*Keywords*)**	**Jenis Skill**
Web Developer	J.620100.01 6.01	Menulis Kode Program	Coding, Programming, Java,PHP, Python, Source Code	Hard Skill
Data Analyst	J.611000.00 2.01	Mengumpulkan Data	Data Collection, Scraping, Query, SQL, Data Mining	Hard Skill
Umum (Semua Jabatan)	LOG.OO01. 002.01	Bekerja efektif dengan orang lain	Teamwork, Kolaborasi, Kerjasama, Komunikasi	Soft Skill

2.3 Metode *Rule-Based*

Pendekatan *Rule-Based* dalam penelitian ini dimanfaatkan untuk mengidentifikasi dan mengklasifikasikan kompetensi berdasarkan basis pengetahuan (*knowledge base*) yang telah ditetapkan sebelumnya. Sistem bekerja menggunakan seperangkat aturan logis yang dirancang untuk mencocokkan kata kunci dalam teks iklan kerja dengan kamus kompetensi *hard skill* dan *soft skill*.Prosesnya dimulai dengan ekstraksi data teks yang kemudian dianalisis terhadap basis aturan yang terstruktur. Keputusan klasifikasi yang dihasilkan mengacu pada logika yang deterministik, jika suatu istilah dalam iklan kerja ditemukan dalam basis pengetahuan, maka sistem akan memberikan label sesuai kategori aturan tersebut. Dengan metode ini, keputusan yang diambil mengikuti pola dan aturan yang sudah dirumuskan secara sistematis.Dalam implementasinya pada klasifikasi teks iklan, metode ini memiliki kriteria penghentian (*stopping condition*) [9]. Penggunaan *Rule-Based* diposisikan sebagai filter pertama untuk menangani pencocokan kata yang bersifat

Copyright © Jurnal Nasional Komputasi dan Teknologi Informasi (JNKTI)

Jurnal Nasional Komputasi dan Teknologi Informasi (JNKTI)
Vol. xx No. xx, mm yyyy
P-ISSN 2620-8342
E-ISSN 2621-3052

pasti (*exact match*), sebelum kemudian dilanjutkan dengan metode *Similarity Matching* untuk menangani variasi istilah yang tidak terdaftar dalam basis pengetahuan.

2.3.1 Pembangunan Basis Pengetahuan (Knowledge Base)

Tahap awal dalam penelitian ini adalah pembangunan basis pengetahuan melalui penyusunan kamus kompetensi yang direpresentasikan dalam struktur data berupa *dictionary*. Basis pengetahuan tersebut memuat kumpulan istilah *hard skill* yang berkaitan dengan kemampuan teknis di bidang profesional IT, serta *soft skill* yang mencerminkan kemampuan non-teknis atau aspek perilaku individu.

2.3.2 Penentuan Aturan Ekstraksi (*Extraction Rules*)

Tahap penentuan aturan ekstraksi dilakukan dengan merumuskan aturan logis yang digunakan sistem untuk mengidentifikasi kandidat kompetensi pada teks iklan lowongan kerja. Proses ini mengacu pada dua pola utama yang menjadi dasar dalam ekstraksi informasi.

a. Aturan Kata Kunci (*Keyword Trigger*): Mendeteksi istilah yang muncul setelah kata pemicu tertentu seperti "menguasai" atau "kualifikasi".

b. Aturan Pola Daftar (*List Pattern*): Mengenali penggunaan simbol daftar (*bullet points*) sebagai penanda bahwa baris teks tersebut memuat poin-poin keahlian.

2.3.3 Mekanisme Pencocokan Bertahap (*Multi-level Matching*)

Proses pemindaian dilakukan secara sistematis untuk menjaga keutuhan makna istilah melalui tingkatan berikut:

a. Pencocokan frasa (*multi-word match*): dilakukan dengan memprioritaskan pencarian istilah yang terdiri atas dua kata atau lebih, seperti “machine learning”, guna mempertahankan keutuhan makna istilah tersebut. b. Pencocokan kata tunggal (*single-word match*): diterapkan apabila frasa yang dimaksud tidak ditemukan, dengan cara melakukan pemeriksaan terhadap setiap token kata secara individual, seperti “java” atau “sql”.

2.3.4 Pelabelan Data (*Data Labeling*)

Tahap pelabelan data (*data labeling*) dilakukan dengan memberikan kategori pada setiap kompetensi yang berhasil diidentifikasi secara otomatis berdasarkan acuan standar. Istilah yang bersifat teknis diklasifikasikan sebagai *hard skill* dengan label numerik, sedangkan kemampuan yang berkaitan dengan aspek interpersonal dikategorikan sebagai *soft skill* dengan label numerik.

2.3.5 Penyaringan Data Terpeta (*Unmapped Filtering*)

Hasil ekstraksi yang memiliki kecocokan secara tepat (*exact match*) dengan basis pengetahuan akan langsung diklasifikasikan ke dalam kategori kompetensi yang sesuai. Adapun istilah yang tidak teridentifikasi dalam basis aturan akan ditandai sebagai data *unmapped* dan selanjutnya diproses pada tahap *similarity matching*.

2.4. Similarity Matching

Bagian ini menjelaskan algoritma yang digunakan untuk mengukur tingkat kemiripan antar data. Adapun algoritma yang digunakan dalam penelitian ini adalah sebagai berikut:

2.4.1 Cosine Similarity

Vektorisasi biner atau pembobotan biner merupakan salah satu teknik paling sederhana dalam merepresentasikan data teks ke dalam bentuk numerik pada bidang pengolahan bahasa alami. Metode ini dilakukan dengan cara mengidentifikasi ada atau tidaknya suatu kata (*term*) dalam sebuah dokumen, tanpa mempertimbangkan jumlah kemunculan kata tersebut. Dalam penerapannya, setiap kata yang termasuk dalam kosa kata (*vocabulary*) yang telah ditentukan akan direpresentasikan sebagai elemen pada sebuah vektor. Apabila suatu kata muncul di dalam dokumen, maka elemen tersebut diberi nilai 1 (*presence*), sedangkan jika tidak muncul akan diberi nilai 0 (*absence*) [10].

Penggunaan vektorisasi biner dalam penelitian ini dinilai tepat untuk proses pencocokan kompetensi (*skill matching*). Hal ini disebabkan karena analisis yang dilakukan berfokus pada kesesuaian antara kualifikasi yang dibutuhkan oleh industri dengan unit kompetensi yang terdapat dalam standar SKKNI. Dengan demikian, keberadaan suatu istilah kunci sudah cukup untuk menunjukkan relevansi kategori, tanpa perlu memperhatikan frekuensi kemunculan kata tersebut dalam teks lowongan kerja.

2.5 Tahapan Text Mining (*Text Preprocessing*)

Proses pengolahan teks bertujuan untuk mengubah dokumen tidak terstruktur menjadi data yang terstruktur sesuai dengan kebutuhan, sehingga dapat diproses lebih lanjut dalam tahapan *text mining*. Tahap pra-pemrosesan teks pada proses klasifikasi berperan penting dalam meningkatkan kualitas data dan memaksimalkan tingkat akurasi hasil klasifikasi data [11].

Text Preprocessing (pra-pemrosesan teks) adalah serangkaian tahapan yang dilakukan secara sistematis untuk membersihkan, menstandarkan, dan menormalkan data teks sebelum digunakan pada proses analisis lebih lanjut. Adapun tahapan pada pra-proses teks antara lain:

1. Data cleaning, Yang bertujuan untuk membersihkan data teks iklan kerja dari *noise* seperti karakter khusus, angka yang tidak relevan, dan duplikasi data.

Copyright © Jurnal Nasional Komputasi dan Teknologi Informasi (JNKTI)

Jurnal Nasional Komputasi dan Teknologi Informasi (JNKTI)
Vol. xx No. xx, mm yyyy
P-ISSN 2620-8342
E-ISSN 2621-3052

2. Case folding, adalah proses mengubah semua huruf dalam teks menjadi huruf kecil sekaligus menghapus tanda baca dan angka. Prosedur ini hanya mempertahankan huruf alfabet dari "a" hingga "z", sedangkan karakter lain selain huruf akan dihilangkan.
3. Tokenizing, Tahapan Ini adalah proses memecah kalimat menjadi kata-kata atau memisahkan deretan karakter dalam sebuah string menjadi potongan-potongan, sehingga setiap kata dalam kalimat dapat dianalisis secara terpisah.
4. Stopword removal, Merupakan tahapan dalam *text preprocessing* yang bertujuan untuk menghilangkan kata-kata umum yang tidak memiliki makna penting atau kontribusi signifikan terhadap isi dokumen, seperti kata penghubung, kata depan, dan kata keterangan.
5. Stemming, Tahapan ini bertujuan untuk memperoleh kata dasar dengan cara menghapus awalan, akhiran, sisipan, serta *confixes* (gabungan awalan dan akhiran) pada kata. Dalam penelitian ini, metode stemming yang digunakan adalah algoritma Snowball Stemmer.

2.6 Evaluasi Performa Confusion Matrix

Confusion matrix, atau yang juga dikenal sebagai *error matrix*, adalah metode yang sering digunakan dalam *machine learning* dan statistik untuk mengevaluasi kinerja suatu model klasifikasi [12]. Matriks ini berfungsi untuk mengevaluasi kinerja model dalam mengklasifikasikan *hard skill* dan *soft skill* pada iklan kerja IT dengan membandingkan hasil prediksi terhadap label sebenarnya, sehingga dapat dihitung beberapa metrik sebagai berikut:

a. Akurasi

Metrik ini mengukur persentase keseluruhan prediksi yang berhasil diklasifikasikan dengan benar (baik *hard skill* maupun *soft skill*) dibandingkan dengan seluruh data iklan kerja IT. Akurasi memberikan gambaran umum tentang seberapa tepat model dalam membedakan dan menandai semua *skill*.

b. Presisi

Precision merupakan metrik evaluasi yang digunakan untuk mengukur tingkat ketepatan sistem dalam menetapkan suatu data *skill* ke dalam kategori positif (*hard skill*).

c. Recall

Recall merupakan metrik evaluasi yang digunakan untuk mengukur kemampuan suatu model dalam mengenali seluruh data yang termasuk ke dalam kategori positif (*hard skill*) secara benar.

d. F1-Score

F1-Score merupakan metrik evaluasi yang menggabungkan nilai *precision* dan *recall* ke dalam satu ukuran untuk menilai kinerja model klasifikasi, terutama pada kondisi data dengan distribusi kelas yang tidak seimbang dalam proses klasifikasi.

2.7 Kebutuhan Perangkat Lunak

2.7.1 Google Colab

Google Colab adalah platform *cloud computing* dari Google yang memungkinkan pengguna menjalankan kode Python secara langsung tanpa instalasi lokal. Platform ini banyak digunakan untuk pemrosesan data, pengembangan kecerdasan buatan, dan pelatihan model mesin [13].

2.7.2 Python

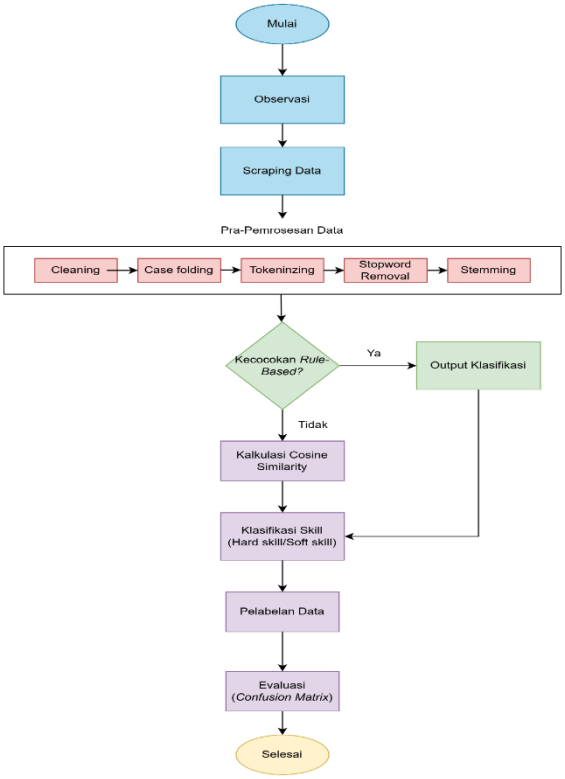
Python adalah bahasa pemrograman interpretatif yang dirancang dengan penekanan pada kejelasan dan kemudahan pemahaman kode. Bahasa ini dibuat untuk mempermudah *programmer* mengembangkan program secara efisien, cepat, dan kompatibel dengan berbagai sistem[14].

3. Objek Penelitian

Objek dalam penelitian ini merupakan kompetensi di bidang teknologi informasi yang mencakup *hard skill* dan *soft skill* pada teks iklan lowongan pekerjaan dari situs Jobstreet. Penelitian ini berfokus pada proses ekstraksi dan klasifikasi istilah kualifikasi yang awalnya berbentuk data teks tidak terstruktur menjadi kategori kompetensi yang lebih sistematis.

Melalui tahapan tersebut, berbagai kebutuhan kompetensi dapat diidentifikasi dan disusun menjadi poin-poin yang lebih terarah. Dengan demikian, hasil klasifikasi memberikan gambaran yang lebih jelas mengenai kualifikasi teknis maupun kemampuan personal yang dibutuhkan oleh perusahaan.

Copyright © Jurnal Nasional Komputasi dan Teknologi Informasi (JNKTI)



Gambar 2.Alur Penelitian

4. HASIL PENELITIAN

4.1 Sumber dan Pengumpulan Data

Data penelitian diperoleh dari platform JobStreet Indonesia, Standar Kompetensi Kerja Nasional Indonesia (SKKNI), dan Peta Okupasi Nasional Indonesia menggunakan teknik *web scraping* dengan bahasa pemrograman Python. Data lowongan pekerjaan digunakan untuk mengidentifikasi kebutuhan keterampilan di bidang IT, sementara data SKKNI digunakan sebagai referensi pengelompokan *hard skill* dan *soft skill*. Pengambilan data pada JobStreet Indonesia dilakukan pada 17 April 2026 dengan perolehan 5.201 data mentah. Setelah melalui tahap *data cleaning* untuk menghapus data duplikat dan tidak relevan, diperoleh 701 data yang valid sebagai dataset utama. Selain itu, diperoleh 2.126 data unit kompetensi dari SKKNI dan 269 data jabatan dari Peta Okupasi melalui proses *scraping* pada tanggal yang sama. Data dari Peta Okupasi digunakan sebagai filter untuk menyaring jabatan pada data lowongan pekerjaan sehingga hanya jabatan yang relevan dengan bidang Teknologi Informasi yang digunakan dalam tahap analisis.

4.2 Seleksi Data

Kumpulan data iklan lowongan kerja dari proses *web scraping* di portal Jobstreet terlebih dahulu melewati proses seleksi awal. Seleksi ini berfokus pada pembersihan data dengan cara menghapus deskripsi kualifikasi yang identik, terduplikasi, atau tidak memiliki informasi lengkap. Proses ini penting dilakukan agar dataset yang digunakan benar-benar bersih, tidak redundan, dan berkualitas baik sebelum masuk ke tahap pengolahan selanjutnya.

4.3 Pra-pemrosesan Data(Text Pre-processing)

4.3.1 Data Cleaning

Tahap *text cleaning* berfungsi untuk membersihkan kolom *Skills* dan *Duties* dari elemen yang tidak diperlukan menggunakan pola *Regular Expression* (re) di Python. Karakter titik (.) dan garis miring (/) tetap dipertahankan karena penting dalam istilah teknis TI, dan hasilnya disimpan dalam kolom *Skills_Clean* dan *Duties_Clean*.

4.3.2 Case Folding

Tahap *case folding* berfungsi mengubah seluruh huruf pada kolom *Skills_Clean* dan *Duties_Clean* menjadi huruf kecil menggunakan metode .str.lower(). Proses ini meminimalkan kesalahan saat proses pencocokan dengan memperlakukan variasi penulisan kata sebagai bentuk yang sama.

Copyright © Jurnal Nasional Komputasi dan Teknologi Informasi (JNKTI)

Jurnal Nasional Komputasi dan Teknologi Informasi (JNKTI)
Vol. xx No. xx, mm yyyy
P-ISSN 2620-8342
E-ISSN 2621-3052

4.3.3 Tokenizing

Tahap *tokenizing* digunakan untuk memecah isi kolom menjadi kata-kata tunggal (token) dengan metode .split() di Python. Teks yang semula berbentuk kalimat diubah menjadi kumpulan kata dalam bentuk daftar (*list*) dan disimpan ke dalam kolom *Skills_Token* dan *Duties_Token*.

4.3.4 Filtering (Stopword Removal)

Tahap *filtering* bertujuan membuang kata-kata umum yang kurang memiliki makna penting menggunakan sumber dari Sastrawi, NLTK, serta daftar tambahan khusus iklan lowongan kerja. Hasilnya disimpan pada kolom *Skills_Filtered* dan *Duties_Filtered* sehingga tersisa kata-kata yang lebih relevan.

4.3.5 Stemming

Tahap *stemming* menghilangkan imbuhan pada setiap token untuk mengembalikan kata ke bentuk dasarnya menggunakan pustaka Sastrawi. Proses ini menggunakan daftar perlindungan untuk istilah teknis TI dan kata berbahasa Inggris agar tidak berubah, kemudian hasilnya disimpan pada kolom *Skills_Stemmed* dan *Duties_Stemmed*.

Seluruh proses *preprocessing* ini diterapkan pada 701 baris dataset iklan lowongan kerja dari Jobstreet yang telah diperoleh melalui proses *web scraping* dan dilanjutkan dengan pembersihan data awal. Dengan demikian, data yang digunakan pada tahap selanjutnya telah berada dalam kondisi yang lebih siap untuk dianalisis.

4.4 Hasil Ekstraksi Kompetensi dengan Metode Rule-Based

Tahap ekstraksi kompetensi berfungsi untuk menyaring hasil dari proses *stemming* sehingga diperoleh daftar keahlian yang benar-benar relevan menggunakan metode *rule-based*. Pendekatan yang digunakan mengacu pada pembangunan *knowledge base* yang berisi daftar kompetensi bidang Teknologi Informasi dan dikelompokkan ke dalam kategori *hard skill* dan *soft skill*.

1. **Kamus *hard skill*** memuat istilah teknis seperti bahasa pemrograman, *framework*, basis data, dan teknologi *cloud*.
2. **Kamus *soft skill*** berisi kemampuan non-teknis seperti komunikasi, kerja tim, dan kepemimpinan.

Untuk menghindari tumpang tindih, sistem menerapkan aturan ekstraksi secara bertahap. Tahap pertama adalah pencocokan frasa (*multi-word match check*) untuk memprioritaskan istilah yang terdiri dari dua kata atau lebih agar makna tetap terjaga. Selanjutnya, sistem melakukan pencocokan kata tunggal (*single-word match check*) di mana setiap token diperiksa satu per satu sesuai entri dalam *knowledge base*. Dengan tahapan ini, proses ekstraksi menghasilkan daftar keahlian yang lebih terarah dan sesuai dengan kebutuhan analisis.

Berdasarkan hasil pembangunan basis pengetahuan (knowledge base), Sistem berhasil mengumpulkan total 1.012 istilah unik dari SKKNI dan Peta Okupasi Nasional, yang terdiri dari kategori *hard skill* (teknis) dan *soft skill* (umum serta etika kerja Islami). Seluruh istilah telah melalui proses *preprocessing* untuk memastikan hanya kata kerja aksi dan objek teknis relevan yang tersimpan. Kedalaman data ini mencakup variasi diksi *bilingual* untuk memperkuat kemampuan algoritma *Cosine Similarity* dalam mendeteksi kemiripan kompetensi secara presisi antara iklan lowongan kerja dengan dokumen formal standar kompetensi.

4.4.1 Pemetaan Peta Okupasi

Setelah memperoleh daftar kompetensi, tahap selanjutnya adalah melakukan pemetaan terhadap standar profesi nasional untuk memastikan kompetensi yang dihasilkan sesuai dengan kebutuhan dunia kerja. Proses pemetaan diawali dengan mencocokkan informasi posisi pekerjaan (*job title*) dari Jobstreet dengan referensi pada Peta Okupasi Nasional TIK melalui penyelarasan istilah pada kolom "Posisi" dan "Judul Okupasi". Untuk meningkatkan tingkat kecocokan, dilakukan penyederhanaan atau normalisasi istilah seperti menghilangkan kata tambahan dan menyesuaikan variasi penulisan jabatan. Hasil dari proses ini berupa pasangan antara data lowongan kerja dengan kategori okupasi yang sesuai, sehingga setiap data dapat dikelompokkan ke dalam klasifikasi pekerjaan yang lebih terstruktur sesuai dengan standar nasional.

4.4.2 Pemetaan Ke Standar Kompetensi Kerja Nasional Indonesia (SKKNI)

Sebagai tahap validasi, hasil pemetaan okupasi dibandingkan dengan dokumen SKKNI pada sektor Teknologi Informasi dan Komunikasi untuk menelusuri unit-unit kompetensi yang relevan. Proses ini memungkinkan keterkaitan antara kompetensi hasil ekstraksi dengan standar resmi nasional, sehingga menunjukkan bahwa metode yang digunakan mampu menghasilkan informasi yang valid, akurat, serta dapat dipertanggungjawabkan.

Pada tahap awal, proses pemindaian menggunakan pendekatan *Rule-Based* terhadap 701 data deskripsi lowongan kerja dengan *knowledge base* yang terdiri dari 173 parameter *hard skill* dan 80 parameter *soft skill*. Berhasil diekstraksi sebanyak 4.868 entri kompetensi tanpa duplikasi, di mana kebutuhan industri didominasi oleh *hard skill* sebesar 62,0% dan *soft skill* sebesar 38,0%.

Hasil ekstraksi tersebut kemudian dipetakan ke dalam Peta Okupasi Nasional bidang TIK dan menemukan 57 jenis okupasi yang relevan, dengan posisi Kepala Portofolio Proyek TI sebagai yang paling dominan sebesar 40,1%. Selanjutnya, dilakukan pemetaan lanjutan ke dalam SKKNI yang menunjukkan bahwa 176 data

Copyright © Jurnal Nasional Komputasi dan Teknologi Informasi (JNKTI)

lowongan kerja memiliki kesesuaian jelas dengan unit kompetensi resmi. Hal ini mengindikasikan bahwa hasil ekstraksi telah selaras dengan kebutuhan kompetensi yang diakui untuk mengukur tingkat akurasi model.

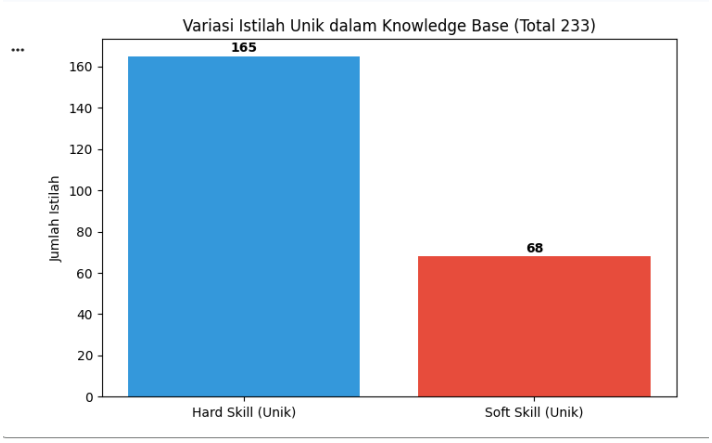
4.4.3 Hasil Pelabelan Data

Tahap pelabelan data dilakukan untuk menyusun acuan nilai yang diproses secara numerik berupa matriks biner, di mana setiap kompetensi diberi nilai Y = 1 untuk kategori *hard skill* dan Y = 0 untuk kategori *soft skill*. Hasil pelabelan ini digunakan sebagai data acuan (*ground truth*) dalam proses evaluasi.

Hasil analisis menunjukkan bahwa kategori *hard skill* memiliki jumlah variasi istilah yang lebih banyak, yaitu sebanyak 165 istilah, dibandingkan dengan *soft skill* yang mencakup 68 istilah. Kondisi ini mencerminkan bahwa bidang Teknologi Informasi menuntut kompetensi teknis yang sangat beragam, mulai dari penguasaan bahasa pemrograman hingga kemampuan dalam menggunakan perangkat lunak desain.

Sebanyak 233 istilah unik yang telah diidentifikasi berperan sebagai dasar utama dalam proses klasifikasi hibrida. Dominasi *hard skill* yang mencapai 62,0% menunjukkan bahwa kebutuhan teknis pada platform Jobstreet bersifat spesifik dan tersegmentasi, sesuai dengan ekosistem teknologi yang digunakan.

Sementara itu, variasi istilah pada *soft skill* cenderung lebih terbatas dan terpusat pada kompetensi interpersonal yang bersifat umum seperti *adaptability*, *analytical thinking*, dan *collaboration*. Dengan demikian, dapat disimpulkan bahwa keragaman *hard skill* menjadi tantangan utama bagi pencari kerja, sedangkan *soft skill* berperan sebagai kompetensi pendukung yang bersifat universal di berbagai posisi IT.



Gambar Grafik Distribusi Istilah Unik pada Basis Pengetahuan(Knowledge Base)

4.5 Hasil Ekstraksi Kompetensi Dengan Metode Similarity Matching

Pada tahap pengolahan data lanjutan, dilakukan proses tambahan untuk mengatasi keterbatasan metode *Rule-Based* dalam mengenali variasi penulisan dan istilah tidak baku yang mengakibatkan sebagian data keterampilan tidak terdeteksi. Untuk mengatasinya, digunakan metode *Cosine Similarity* yang berfungsi mengukur tingkat kemiripan antar teks berdasarkan konteks, bukan hanya kesamaan kata.

Dengan pendekatan ini, istilah yang memiliki makna serupa meskipun berbeda penulisan tetap dapat dikenali dan diklasifikasikan dengan tepat. Penerapan metode ini membantu meningkatkan cakupan identifikasi kompetensi serta memperbaiki hasil klasifikasi secara keseluruhan.

4.5.1 Cosine Similarity Matching

Dalam pemetaan kualifikasi tenaga kerja IT dari iklan lowongan kerja, sering dijumpai penulisan frasa tidak baku dan singkatan yang tidak dapat dikenali secara tepat oleh metode *rule-based*. Sebagai solusi, penelitian ini menerapkan algoritma *Cosine Similarity* untuk mengidentifikasi kualifikasi yang tidak berhasil dikenali pada tahap sebelumnya dengan mengukur tingkat kemiripan antara data lowongan kerja dengan data acuan pada tabel *ground truth*.

Sebelum perhitungan dilakukan, data teks ditransformasikan ke bentuk numerik menggunakan teknik vektorisasi biner, di mana setiap istilah direpresentasikan sebagai elemen dalam sebuah vektor. Pendekatan ini memungkinkan representasi data yang seragam sehingga dapat langsung diproses menggunakan *Cosine Similarity* untuk menghitung skor kemiripan. Melalui kombinasi ini, sistem mampu mengakomodasi perbedaan istilah yang memiliki makna serupa sehingga proses identifikasi kualifikasi menjadi lebih fleksibel dan komprehensif.

4.5.1.1 Tahapan Vektorisasi Teks Biner (Bag of Words)

Data teks pada deskripsi lowongan diubah ke dalam bentuk numerik melalui proses vektorisasi menggunakan pendekatan *Bag of Words* (BoW) dengan modul *CountVectorizer*. Setiap kata direpresentasikan

Copyright © Jurnal Nasional Komputasi dan Teknologi Informasi (JNKTI)

Jurnal Nasional Komputasi dan Teknologi Informasi (JNKTI)
Vol. xx No. xx, mm yyyy
P-ISSN 2620-8342
E-ISSN 2621-3052

ke dalam bentuk matriks biner (0 dan 1) berdasarkan keberadaannya dalam teks untuk menentukan kesamaan token dengan unit kompetensi pada korpus SKKNI.

4.5.1.2 Perhitungan Kemiripan Sudut (Dot Product)

Setelah data direpresentasikan dalam matriks biner, dilakukan perhitungan tingkat kemiripan menggunakan algoritma *Cosine Similarity* melalui perkalian elemen vektor (*dot product*). Apabila sudut antar kedua vektor semakin kecil, maka nilai *cosine* akan mendekati 1, yang menunjukkan bahwa kedua teks memiliki tingkat kesamaan yang tinggi atau relevansi yang kuat.

4.5.1.3 Batas Ambang Kelayakan (Threshold Limit)

Hasil prediksi skor kemiripan dievaluasi menggunakan filter ambang batas (*threshold*) dengan toleransi minimum sebesar ≥ 0,1. Rekaman kualifikasi yang berhasil menyeberangi margin tersebut dinyatakan valid dan dicap (*labeled*) ke dalam kelompok *hard skill* atau *soft skill*, sedangkan skor di bawah 0,1 akan dinyatakan sebagai data melenceng (*noise*) untuk digugurkan.

4.5.1.4 Ilustrasi Rekapitulasi Penilaian Metrik

Bagian ini menyajikan visualisasi perbandingan teks menggunakan metode *Cosine Similarity*. Pada data dengan korelasi tinggi, sistem mampu mengidentifikasi kesesuaian antara teks pelamar dengan acuan secara tepat, sedangkan pada data dengan nilai kemiripan rendah, sistem tidak menganggapnya relevan sehingga data tersebut tidak dipertahankan dalam proses selanjutnya.

4.5.1.5 Ilustrasi Perhitungan Manual Cosine Similarity

Pada tahap ini dilakukan pembuktian perhitungan manual terhadap metode *Cosine Similarity* dengan merepresentasikan data dalam bentuk vektor untuk menghitung nilai kemiripan pada rentang 0 hingga 1. Penggunaan metode ini didasarkan pada temuan bahwa terdapat 143 data dengan penulisan yang tidak konsisten atau kesalahan ejaan yang tidak mampu dideteksi secara optimal oleh metode *rule-based*. Melalui perhitungan manual, terlihat bahwa *Cosine Similarity* tetap mampu mengidentifikasi kemiripan antar data meskipun terdapat perbedaan penulisan, sehingga metode ini digunakan untuk mengatasi keterbatasan pendekatan *rule-based*.

4.5.1.6 Rekapitulasi Hasil Pemrosesan Data

Algoritma *Cosine Similarity* dengan pendekatan vektorisasi biner mampu mengolah sebanyak 143 data kualifikasi pelamar (*unmapped data*) yang sebelumnya tidak berhasil dipetakan. Dari proses pengolahan tersebut, diperoleh ringkasan hasil sebagai berikut:

a. **Sebanyak 141 data** kualifikasi berhasil teridentifikasi dan diklasifikasikan sebagai Lulus (Valid) karena memiliki nilai *Cosine Similarity* ≥ 0,1.
b. **Sebanyak 2 data** dikategorikan sebagai Gagal (Bukan Skill) karena nilai kemiripan berada di bawah batas minimum.

Terhadap 141 data yang dinyatakan lulus, dilakukan proses pelabelan otomatis dengan mengacu pada kategori unit kompetensi SKKNI yang memiliki tingkat kemiripan tertinggi. Hasil pelabelan menunjukkan bahwa keterampilan yang teridentifikasi terdiri dari 110 data kategori *hard skill* dan 33 data kategori *soft skill*.

Tabel Sampel data Perhitungan Manual

No.	Vektor A (Jobstreet)	Vektor B (SKKNI)	Vektor B (SKKNI)	A*B
1.	510	1	0	0
2.	a	1	0	0
3.	across	1	0	0
4.	analytical	1	0	0
5.	and	1	0	0
6	applications	1	0	0
7	bachelors	1	0	0
8.	career	1	0	0
9	collaborative	1	0	0
10	commercial	1	0	0
11	communication	1	1	1
12	comply	1	0	0

Copyright © Jurnal Nasional Komputasi dan Teknologi Informasi (JNKTI)

Jurnal Nasional Komputasi dan Teknologi Informasi (JNKTI)
Vol. xx No. xx, mm yyyy
P-ISSN 2620-8342
E-ISSN 2621-3052

12	dan	1	0	0
14	degree	1	0	0
15	development	1	0	0
16	driving	1	0	0
17	dynamic	1	0	0
18	engineering	1	1	1
19	english	1	0	0
20	environment	1	0	0
21	experience	1	0	0
22	exposure	1	0	0
23	extensively	1	0	0
24	field	1	0	0
25	fleet	1	0	0
26	fluent	1	0	0
27	for	1	0	0
28	good	1	0	0
29	growth	1	0	0
30	in	1	0	0
31	indonesia	1	0	0
32	industry	1	0	0
33	intelektual	0	1	0
34	intellectual	1	1	1
35	join	1	0	0
36	kekayaan	0	1	0
37	key	1	0	0
38	knowledge	1	0	0
39	leading	1	0	0
40	license	1	0	0
41	mematuhi	0	1	0
42	minimum	1	0	0
43	nationwide	1	0	0
44	of	1	0	0
45	operations	1	0	0
46	opportunities	1	0	0
47	opportunity	1	0	0
48	or	1	0	0
49	organisasi	0	1	0
50	organisational	1	0	0
51	penggunaan	0	1	0

Copyright © Jurnal Nasional Komputasi dan Teknologi Informasi (JNKTI)

Jurnal Nasional Komputasi dan Teknologi Informasi (JNKTI)
Vol. xx No. xx, mm yyyy
P-ISSN 2620-8342
E-ISSN 2621-3052

52	perlindungan	0	1	0
53	persyaratan	0	1	0
54	players	1	0	0
55	possess	1	0	0
56	preferred	1	0	0
57	presentation	1	0	0
58	products	1	0	0
59	property	1	1	1
60	protection	1	0	0
61	related	1	0	0
62	requirements	1	0	0
63	sim	0	1	0
64	skills	1	0	0
65	strong	1	0	0
66	tbr	1	0	0
67	the	1	0	0
68	Tire	0	1	0
69	to	1	0	0
70	travel	0	1	0
71	Untuk	0	1	0
72	us	0	1	0
73	use	0	1	0
74	valid	0	1	0
75	vehicles	0	1	0
76	why	1	0	0
77	willing	1	0	0
78	with	1	0	0
79	work	1	0	0
80	working	1	0	0
81	years	1	0	0
		∑(A)=65	∑(B)=20	∑(A*B)=4

Perhitungan kemiripan antara vektor biner hasil *Vektorisasi Biner* dilakukan menggunakan algoritma *Cosine Similarity* dengan rumus sebagai berikut:

$$similarity = cos(\theta) = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

1. Hitung Dot Product (A*B)
   $\sum(A_i * B_i)$= 4

Copyright © Jurnal Nasional Komputasi dan Teknologi Informasi (JNKTI)

Jurnal Nasional Komputasi dan Teknologi Informasi (JNKTI)
Vol. xx No. xx, mm yyyy
P-ISSN 2620-8342
E-ISSN 2621-3052

$$A.B = (1 \times 0) + (1 \times 0) + (1 \times 0) + (1 \times 0) + (1 \times 0) + (1 \times 0) + (1 \times 0) + (1 \times 0) + (1 \times 0) + (1 \times 1) + (1 \times 0) + (1 \times 0) + (1 \times 0) + (1 \times 0) + (1 \times 0) + (1 \times 0) + (1 \times 1) + (1 \times 0) + (1 \times 0) + (1 \times 0) + (1 \times 0) + (1 \times 0) + (1 \times 0) + (1 \times 0) + (1 \times 0) + (1 \times 0) + (1 \times 0) + (1 \times 0) + (1 \times 0) + (0 \times 1) + (1 \times 0) + (1 \times 0) + (1 \times 0) + (1 \times 0) + (1 \times 0) + (1 \times 0) + (1 \times 0) + (1 \times 0) + (0 \times 1) + (1 \times 0) + (0 \times 1) + (0 \times 1) + (0 \times 1) + (1 \times 0) + (1 \times 0) + (1 \times 0) + (1 \times 0) + (1 \times 0) + (1 \times 1) + (1 \times 0) + (1 \times 0) + (1 \times 0) + (0 \times 1) + (1 \times 0) + (1 \times 0) + (1 \times 0) + (1 \times 0) + (0 \times 1) + (1 \times 0) + (0 \times 1) + (0 \times 1) + (0 \times 1) + (0 \times 1) + (0 \times 1) + (1 \times 0) + (1 \times 0) + (1 \times 0) + (1 \times 0) = 4$$

2. Hitung Magnitudo Vektor

$\|A\| = \sqrt{\sum_{i=1}^{n} A_i^2}$ $\|A\| = \sqrt{65} = 8{,}062$

$\|B\| = \sqrt{\sum_{i=1}^{n} B_i^2}$ $\|B\| = \sqrt{20} = 4{,}472$

3. Hitung Cosine Simalirity

$$Cosine\ Similarity\ (A, B) = \frac{\text{A. B}}{\|A\| \times \|B\|}$$

$$Cosine\ Similarity\ (A, B) = \frac{4}{\sqrt{8{,}062 \times\ 4.472}} = \frac{4}{36.053} = 0.1109$$

Berdasarkan hasil perhitungan manual yang telah dilakukan, diperoleh nilai kemiripan sebesar (> 0.1). Nilai tersebut mengindikasikan adanya tingkat kesesuaian yang sangat tinggi antara kualifikasi yang dibutuhkan industri dengan standar SKKNI.

Karena nilai kemiripan tersebut melampaui ambang batas (*threshold*) yang ditetapkan, maka sistem secara otomatis mengklasifikasikan kualifikasi tersebut sebagai data yang valid serta memetakan istilah "kemahiran menulis kode baris bahasa pemrograman" ke dalam unit kompetensi "menulis kode baris program". Penerapan pendekatan hibrida ini menunjukkan kemampuan dalam mengakomodasi variasi penggunaan bahasa alami pada iklan lowongan kerja, tanpa mengurangi keselarasan terhadap standar kompetensi nasional yang berlaku.

4.5 Hasil Klasifikasi Akhir (Penentuan Kategori)

Setelah seluruh tahapan pemrosesan dilakukan, mulai dari penerapan metode berbasis aturan (*rule-based*) hingga penanganan data melalui pendekatan perhitungan kemiripan vektor (*Cosine Similarity*), penelitian ini menghasilkan klasifikasi akhir yang lebih menyeluruh.

Kedua metode tersebut digunakan secara terpadu sehingga saling melengkapi, di mana data yang tidak memiliki kesesuaian ejaan secara langsung dengan kamus referensi tetap dapat diklasifikasikan melalui pendekatan matematis. Hasil penggabungan klasifikasi terhadap seluruh istilah kompetensi yang telah diekstraksi kemudian disajikan pada tabel berikut:

Tabel 2. Jumlah Data Aktual

No.	Kategori Kompetensi	Label Sistem (Y)	Jumlah Data
1.	*Hard Skill*	1	519
2.	*Soft Skill*	0	182
3.	Total Kompetensi Berhasil Diklasifikasikan		701

4.6 Evaluasi Performa Model (Confusion Matrix)

Tahap evaluasi dilakukan untuk menilai sejauh mana efektivitas algoritma *Cosine Similarity* dalam mengklasifikasikan kualifikasi pekerjaan menggunakan metode *Confusion Matrix*. Berdasarkan hasil pengujian tersebut, diperoleh empat metrik utama untuk menilai kinerja model secara keseluruhan:

a. **Akurasi (*Accuracy*):** Nilai akurasi sebesar 88,02% menunjukkan tingkat ketepatan sistem dalam melakukan klasifikasi secara keseluruhan. Hal ini mengindikasikan bahwa penerapan metode hibrida mampu mengatasi perbedaan istilah antara deskripsi lowongan kerja dan standar SKKNI dengan cukup efektif.

Copyright © Jurnal Nasional Komputasi dan Teknologi Informasi (JNKTI)

Jurnal Nasional Komputasi dan Teknologi Informasi (JNKTI)
Vol. xx No. xx, mm yyyy
P-ISSN 2620-8342
E-ISSN 2621-3052

b. **Presisi (*Precision*):** Nilai presisi sebesar 88,09% mencerminkan tingkat ketepatan sistem dalam mengidentifikasi kategori *Hard Skill*. Nilai presisi yang tinggi ini menunjukkan bahwa kesalahan klasifikasi dapat diminimalkan, sehingga hasil pemetaan tetap relevan dengan kebutuhan industri.
c. **Recall (Sensitivitas):** Nilai *recall* sebesar 96,92% menunjukkan bahwa sistem memiliki kemampuan yang sangat baik dalam mendeteksi seluruh data yang seharusnya termasuk dalam kategori *Hard Skill*. Keunggulan ini menandakan bahwa sistem efektif dalam mengenali berbagai bentuk penulisan kualifikasi teknis.
d. **F1-Score:** Nilai *F1-Score* sebesar 92,29% menunjukkan bahwa model memiliki kinerja yang stabil serta tidak cenderung bias terhadap salah satu kategori. Oleh karena itu, *F1-Score* dapat menjadi indikator yang representatif dalam menilai kualitas keseluruhan model klasifikasi.

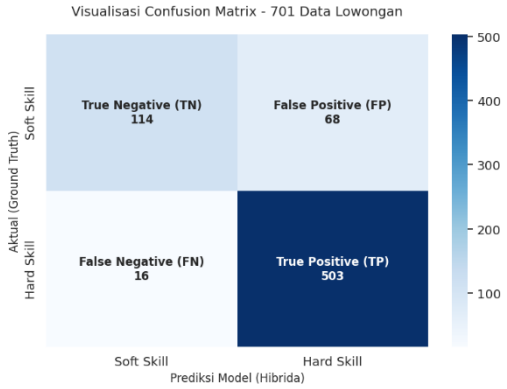
```
============================================================
HASIL EVALUASI PERFORMA MODEL HIBRIDA (701 DATA)
============================================================
Akurasi (Accuracy)   : 88.02%
Presisi (Precision)  : 88.09%
Recall               : 96.92%
F1-Score             : 92.29%
------------------------------------------------------------

Classification Report Lengkap:
              precision    recall  f1-score   support

  Soft Skill       0.88      0.63      0.73       182
  Hard Skill       0.88      0.97      0.92       519

    accuracy                           0.88       701
   macro avg       0.88      0.80      0.83       701
weighted avg       0.88      0.88      0.87       701
```

Gambar berikut menyajikan *confusion matrix* sebagai alat evaluasi performa model klasifikasi dalam membedakan kategori *soft skill* dan *hard skill* yang terdiri atas empat komponen utama:



a. True Positive (TP) = 503: Jumlah data *hard skill* yang berhasil diklasifikasikan dengan benar sebagai *hard skill*, menunjukkan model cukup baik dalam mengenali *hard skill*.
b. True Negative (TN) = 114: Jumlah data *soft skill* yang berhasil diklasifikasikan dengan benar sebagai *soft skill*, mengindikasikan model memiliki kemampuan yang baik dalam mengidentifikasi *soft skill*.
c. False Positive (FP) = 68: Data *soft skill* yang keliru diklasifikasikan sebagai *hard skill*, menunjukkan kecenderungan model menganggap beberapa *soft skill* sebagai *hard skill*.
d. False Negative (FN) = 16: Data *hard skill* yang salah diklasifikasikan sebagai *soft skill*, di mana kesalahan dalam mendeteksi *hard skill* relatif lebih rendah.

Secara keseluruhan, model menghasilkan akurasi sebesar 88,02% yang menunjukkan sebagian besar prediksi sudah sesuai dengan label sebenarnya. Dengan dominasi nilai TP dan TN yang cukup tinggi, dapat disimpulkan bahwa model memiliki performa yang baik dalam melakukan klasifikasi, meskipun masih terdapat sejumlah kesalahan prediksi pada kategori *false positive*.

Berdasarkan Confusion Matrix, dilakukan metrik evaluasi yang meliputi accuracy, precision, recall, dan F1-Score dengan menggunakan rumus sebagai berikut:

$$Accuary = \frac{TP + TN}{TP + TN + FP + FN} = \frac{503 + 114}{503 + 114 + 68 + 16}$$

Copyright © Jurnal Nasional Komputasi dan Teknologi Informasi (JNKTI)

Jurnal Nasional Komputasi dan Teknologi Informasi (JNKTI)
Vol. xx No. xx, mm yyyy
P-ISSN 2620-8342
E-ISSN 2621-3052

$$= \frac{617}{701} = 88{,}02\%$$

$$Precision = \frac{TP}{TP + FP} = \frac{503}{503 + 68} = \frac{503}{571} = 88{,}09\%$$

$$Recall = \frac{TP}{TP + FN} = \frac{503}{503 + 16} = \frac{503}{519} = 96{,}92\%$$

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision \times Recall} = \frac{2 \times (0{,}8809 \times 0{,}9692)}{0{,}8809 + 0{,}9692}$$

$$F1 - Score = \ 2 \times \frac{(0{,}8537)}{(1.8501)}$$

$$= \frac{1.7074}{1.8501}$$

$$= 0{,}9229$$

4.7 Analisis Hasil Evaluasi dan Performa Sistem

Berdasarkan hasil pengujian menggunakan *Confusion Matrix*, diperoleh nilai evaluasi berupa *accuracy*, *precision*, *recall*, dan *F1-score* untuk mengukur kinerja sistem dalam mengklasifikasikan data sesuai standar SKKNI. Secara keseluruhan, hasil menunjukkan bahwa penerapan metode *Rule-Based* yang dikombinasikan dengan *Similarity Matching* mampu mengidentifikasi serta mengelompokkan kompetensi dari teks iklan lowongan kerja yang tidak terstruktur dengan cukup efektif.

4.7.1 Interpretasi Nilai Akurasi

Nilai akurasi menunjukkan persentase jumlah prediksi yang benar dibandingkan dengan keseluruhan data uji. Tingkat akurasi yang baik dipengaruhi oleh penggunaan *knowledge base* berbasis SKKNI sebagai acuan objektif, efektivitas metode *Rule-Based* dalam pencocokan langsung, serta peran *Cosine Similarity* dalam menangani variasi istilah yang memiliki perbedaan penulisan namun mengandung makna yang sama.

4.7.2 Analisis Precision, Recall, dan F1-Score

Nilai *precision* menunjukkan tingkat ketepatan sistem dalam mengklasifikasikan kategori *hard skill*, sementara *recall* menggambarkan kemampuan sistem dalam mengidentifikasi seluruh data yang seharusnya termasuk dalam kategori tersebut. *F1-score* digunakan untuk melihat keseimbangan kinerja sistem secara keseluruhan, yang menunjukkan bahwa sistem memiliki ketepatan klasifikasi sekaligus konsisten mendeteksi data yang relevan.

4.7.3 Perbandingan Performa Rule-Based dan Similarity Matching

Kedua metode yang digunakan memiliki peran yang saling melengkapi; *Rule-Based* menangani istilah yang memiliki kecocokan langsung secara cepat dan akurat, sedangkan *Similarity Matching* berfungsi sebagai pendukung dalam menangani variasi istilah, sinonim, serta penggunaan bahasa yang beragam.

5.Kesimpulan

Berdasarkan penelitian mengenai klasifikasi kompetensi IT dari iklan lowongan kerja Jobstreet, penggunaan pendekatan hibrida yang mengintegrasikan metode *Rule-Based* dan *Similarity Matching* terbukti sangat efektif dalam mengekstraksi serta mengelompokkan talenta digital secara sistematis. Metode *Rule-Based* berhasil mengidentifikasi 4.868 entri kompetensi unik dengan dominasi *hard skill* sebesar 62%, sementara algoritma *Cosine Similarity* memberikan kontribusi signifikan dalam mengatasi variasi penulisan istilah tidak baku sehingga meningkatkan cakupan data yang sebelumnya tidak terdeteksi. Hasil evaluasi menggunakan *confusion matrix* menunjukkan performa sistem yang andal dengan nilai akurasi mencapai 88,02% dan *recall* sebesar 96,92%, yang mengindikasikan bahwa kombinasi kedua metode ini mampu meminimalkan kesalahan identifikasi dan selaras dengan kebutuhan standar kompetensi nasional (SKKNI).

Copyright © Jurnal Nasional Komputasi dan Teknologi Informasi (JNKTI)

[1] R. Vidyasari and R. Listiawati, "Analisis Kebutuhan Hard Skill Komputer untuk Program Studi Sarjana Terapan Keuangan dan Perbankan PNJ," vol. 8, no. 1, pp. 288–299, 2022.
[2] H. Geraldy, R. A. Farentina, and F. A. Dirk, "Identifikasi Keterampilan Digital dalam Iklan Lowongan Kerja Menggunakan Klasifikasi Teks dan Named Entity Recognition Studi Kasus pada Portal Jobstreet," no. 2020, pp. 247–256, 2025.
[3] M. Malik and I. Malang, "Analisis Kesesuaian Lowongan Pekerjaan dan Kemampuan Pencari Kerja dengan Menggunakan Metode Cosine Similarity," vol. 24, no. 3, pp. 617–631, 2025.
[4] HUSNA, "Pengaruh knowledge , skill , dan attitude terhadap kinerja pegawai pada kantor dinas perhubungan kabupaten enrekang," 2022.
[5] Y. R. G. Ratuela1, "Y . R . G . Ratuela ., O . S . Nelwan ., G . G . Lumintang . KERJA PADA MAHASISWA AKHIR JURUSAN MANAJEMEN FEB UNSRAT MANADO THE INFLUENCE OF HARD SKILLS , SOFT SKILLS AND SELF-EFFICACY ON WORK READINESS IN FINAL STUDENTS MAJORING IN MANAGEMENT FEB UNSRAT ," vol. 10, no. 1, pp. 172–183, 2022.
[6] N. Huda, "Jurnal Pengabdian Masyarakat Sains dan Teknologi Volume. 3, No. 4, Tahun 2024," no. 4, 2025.
[7] Kementerian Komunikasi dan Informatika RI., "Peta Okupasi Nasional Bidang Teknologi Informasi dan Komunikasi (TIK).," *Kementeri. Komun. dan Inform. RI*, 2017.
[8] Kementerian Ketenagakerjaan Republik Indonesia, *Keputusan Menteri Ketenagakerjaan Republik Indonesia Nomor 282 Tahun 2016 tentang Standar Kompetensi Kerja Nasional Indonesia Bidang Teknologi Informasi dan Komunikasi*. Jakarta, 2016. [Online]. Available: https://skkni.kemnaker.go.id/dokumen
[9] A. M. Aubaid, A. Mishra, and A. Mishra, "Machine learning and rule-based embedding techniques for classifying text documents," *Int. J. Syst. Assur. Eng. Manag.*, vol. 15, no. 12, pp. 5637–5652, 2024, doi: 10.1007/s13198-024-02555-w.
[10] M. R. Waskito, A. D. Rahajoe, A. L. Nurlaili, and J. Timur, "COLLABORATIVE FILTERING MENGGUNAKAN ALGORITMA COSINE SIMILARITY DAN JACCARD SIMILARITY PADA SISTEM E-COMMERCE," vol. 12, no. 3, pp. 4307–4316, 2024.
[11] T. Ridwansyah, "Implementasi Text Mining Terhadap Analisis Sentimen Masyarakat Dunia Di Twitter Terhadap Kota Medan Menggunakan K-Fold Cross Validation Dan Naïve Bayes Classifier," vol. 2, no. 5, pp. 178–185, 2022.
[12] M. Fadhil, A. Hasril, and L. Budi, "Implementasi Algoritma K-Nearest Neighbors Untuk Klasifikasi Data Ulasan Pengguna Aplikasi Sulselbar Mobile pada Google Play Store," vol. 2, no. 3, pp. 358–366, 2025.
[13] A. Y. Roni Andarsyah1, "Jurnal Teknik Informatika, Vol. 16, No. 2, April 2024," vol. 16, no. 2, pp. 1–7, 2024.
[14] S. J. Maulidiyah and A. I. Syahyadi, "Analysis Of E-Commerce Product With Web Scraping Technique," vol. 3, no. 1, pp. 37–46, 2025.

Copyright © Jurnal Nasional Komputasi dan Teknologi Informasi (JNKTI)