Kelompok: 10
Anggota: - Muhammad Fariz Abid R (23031554030)
- Hasna Lestari (23031554033)
- Arfan Hafidt Nashrullah (23031554112)

Optimalisasi Proses Penyaringan Kandidat Berdasarkan Kecocokan CV dengan Teknologi Digital

A. Latar Belakang

Di era digital yang terus berkembang, perusahaan menghadapi tantangan signifikan dalam proses rekrutmen, khususnya dalam penyaringan Curriculum Vitae (CV) secara manual. Jumlah pelamar yang meningkat pesat membuat tugas ini semakin kompleks dan memakan waktu bagi departemen Sumber Daya Manusia (HRD). Proses manual tidak hanya memerlukan waktu yang lama tetapi juga rentan terhadap kesalahan manusia, yang dapat menyebabkan ketidakcocokan antara kualifikasi kandidat dan kebutuhan perusahaan.

Untuk mengatasi tantangan ini, berbagai inovasi teknologi telah diperkenalkan dalam proses seleksi karyawan. Salah satu solusi yang menonjol adalah penggunaan Sistem Pelacakan Pelamar (Applicant Tracking System/ATS). ATS memungkinkan HRD untuk mengelola dan menyortir aplikasi secara otomatis berdasarkan kriteria tertentu, sehingga mempercepat proses penyaringan dan memastikan kandidat yang sesuai dapat diidentifikasi dengan cepat (Raflika et al., 2024).

Selain itu, teknologi kecerdasan buatan (Artificial Intelligence/AI) telah membawa perubahan signifikan dalam proses rekrutmen. AI dapat membantu dalam mengevaluasi sejumlah besar aplikasi dalam waktu singkat, mengidentifikasi pola perilaku dan keterampilan, serta memberikan analisis yang mendalam. Sebuah studi kasus pada PT Saripetejo Sukabumi menunjukkan bahwa implementasi AI meningkatkan efisiensi proses rekrutmen hingga 60% dengan akurasi pencocokan kandidat sebesar 75%(Wahidin et al., 2024).

Teknologi lain yang mulai diterapkan adalah wawancara video dan asesmen online. Metode ini memungkinkan HRD untuk melakukan wawancara dan penilaian secara virtual, mengurangi kebutuhan pertemuan tatap muka awal, serta memberikan gambaran lebih mendalam tentang keterampilan dan kepribadian calon karyawan (Raflika et al., 2024).

Dengan memanfaatkan teknologi-teknologi tersebut, perusahaan dapat mempercepat dan mempermudah proses rekrutmen tanpa mengorbankan kualitas seleksi. Implementasi solusi berbasis teknologi ini tidak hanya meningkatkan efisiensi tetapi juga membantu HRD dalam membuat keputusan yang lebih akurat dan objektif dalam memilih kandidat yang tepat.

B. Dataset

Data yang digunakan ini terdapat 2484 CV kandidat pekerja yang bisa dilakukan penyaringan untuk mencari kecocokan CV dengan pekerjaan untuk proyek ini yaitu https://www.kaggle.com/datasets/snehaanbhawal/resume-dataset

C. Manfaat

1. **Penyaringan yang Lebih Efisien**: Mempercepat proses pencarian kandidat yang sesuai dengan kriteria perusahaan, memungkinkan HRD untuk fokus pada kandidat yang lebih relevan.
2. **Mengurangi Human Error**: Dengan otomatisasi dalam membaca dan menganalisis CV, risiko kesalahan manusia dalam proses screening dapat diminimalisir.
3. **Penghematan Waktu**: Meningkatkan efisiensi waktu dalam tahap penyaringan awal, memungkinkan HRD untuk mengalokasikan waktu lebih banyak pada tahap seleksi berikutnya.
4. **Peningkatan Akurasi**: Teknologi dapat membantu mengidentifikasi kecocokan antara pengalaman, keterampilan, dan kualifikasi kandidat membantu mengidentifikasi kecocokan antara pengalaman, keterampilan, dan kualifikasi kandidat.

D. Tujuan

1. **Meningkatkan Efisiensi Proses Rekrutmen**: Menciptakan sistem yang dapat mempercepat proses penyaringan CV dan membantu HRD untuk menemukan kandidat yang tepat lebih cepat, sehingga menghemat waktu dan tenaga.
2. **Meminimalisir Human Error dalam Screening**: Mengurangi kesalahan yang terjadi akibat keterbatasan manusia dalam menganalisis dan membandingkan banyaknya CV, memastikan proses seleksi berjalan lebih akurat dan objektif.
3. **Meningkatkan Kualitas Kandidat yang Ditemukan**: Dengan menggunakan teknologi yang dapat menilai kecocokan CV secara lebih mendalam, diharapkan dapat menemukan kandidat yang benar-benar sesuai dengan kebutuhan perusahaan.
4. **Menyederhanakan Proses Seleksi Awal**: Membantu HRD untuk menyaring pelamar secara lebih sistematis, mengurangi beban kerja manual, dan memungkinkan fokus pada kandidat yang memiliki potensi terbaik.
5. **Mengoptimalkan Penggunaan Sumber Daya**: Menggunakan teknologi untuk mengelola proses seleksi dengan lebih hemat biaya dan waktu, sehingga HRD dapat mengalokasikan lebih banyak sumber daya untuk tahap wawancara dan evaluasi kandidat yang lebih mendalam.

E. Pre-Processing

- Remove URL: menghapus URL yang diawali dengan http, https, dan www.
- Remove mention: Menghapus mention yang diawali dengan @ kemudian diikuti username
- Remove hashtag: Menghapus kata-kata yang dimulai dengan tanda #
- Remove Kanji Character dan Unicode: menghapus karakter kanji dan karakter unicode dengan range tertentu
- Remove Character and Punctuation: Menghapus karakter selain huruf
- Remove Numbers: Menghapus angka pada kalimat
- Remove Single Character: Menghapus kata-kaya yang hanya terdiri satu karakter huruf
- Remove Multiple Space: Mengganti beberapa spasi berturut-turut menjadi satu spasi
- Lower Casing: Mengubah semua teks menjadi huruf kecil

- Tokenize, Remove Stopwords, dan Lemmatize: Memecah teks menjadi sebuah kata, menghapus kata yang bersifat umum dan ridak memiliki banyak informasi, dan mengubah sebuah kata menjadi kata dasar sesuai KBBI
- Remove Keyword: Menghapus beberapa kata kunci seperti (company, name, city, state) kemudian diubah menjadi pola regex
- Menghapus Spasi Ekstra: Menghapus spasi yang tidak seharusnya (lebih dari 1)

F. Feature Engineering

a. Teknik Feature Engineering

- TF-IDF (Term Frequency-Inverse Document Frequency):

  Salah satu algoritma ekstraksi ciri yang paling umum digunakan adalah TF-IDF (Cahyanti et al., 2020). Nilai frekuensi kemunculan suatu kata dalam suatu dataset dikenal dengan term frequency (TF) (Hasan et al., n.d.). Sementara itu, Inverse Document Frequency (IDF) digunakan untuk menentukan tingkat kepentingan kata tersebut dalam dataset. Suatu term akan memiliki bobot yang rendah jika sering muncul dalam setiap dokumen dalam dataset; jika tidak, maka akan memiliki bobot yang lebih besar (Ranjan & Mishra, 2020). Rumus untuk menghitung TF-IDF adalah Persamaan dibawah, di mana nilai tf adalah frekuensi kemunculan kata t dalam dokumen word, df adalah banyaknya dokumen yang memuat kata t, dan N adalah banyaknya dokumen dalam dataset.

$$TF - IDF(word) = TF(word) \times log(\frac{N}{df(word)})$$

- Word2Vec

  Pertama kali dikembangkan oleh Mikolov et al., pada tahun 2013, Word2Vec merupakan sebuah metodologi untuk memperoleh vektor kata yang memiliki manfaat untuk dapat menangkap makna semantik dari sebuah kata (Shi et al., 2019) (Muhammad et al., n.d.). Arsitektur Word2Vec didasarkan pada gagasan jaringan saraf dua lapis, dengan data teks sebagai input dan data vektor sebagai output (Nawangsari et al., n.d.). Dapat digunakan untuk membentuk vektor kalimat atau dokumen berdasarkan representasi kata, yang memberikan pemahaman konteks yang lebih dalam daripada TF-IDF.

b. EDA

```
<class 'pandas.core.frame.DataFrame'>
Index: 2483 entries, 0 to 2483
Data columns (total 6 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   ID               2483 non-null   int64
 1   Resume_str       2483 non-null   object
 2   Resume_html      2483 non-null   object
 3   Category         2483 non-null   object
 4   full_text_clean  2483 non-null   object
 5   final_text_clean 2483 non-null   object
dtypes: int64(1), object(5)
memory usage: 135.8+ KB
```

Tipe data merupakan text atau object yang berisi kumpulan CV. Pada setiap CV berisi deskripsi diri, pengalaman, skill serta kemampuan yang dimiliki pelamar.

G. Metode

- **Jaccard Similarity**

Jaccard Similarity adalah sebuah algoritma dari kelas **q-gram** (Gali et al., n.d.). Teknik ini digunakan untuk menentukan tingkat kesamaan antara dua objek (item). Perhitungan dalam pendekatan ini, seperti cosine distance dan matching coefficient, umumnya didasarkan pada kesamaan dalam ruang vektor ukuran (Pikies & Ali, 2019).

Gagasan utama di balik algoritma ini adalah bahwa urutan karakter sama pentingnya dengan karakter itu sendiri (Diana & Hanana Ulfa, 2019). Dalam dua dokumen, metode Jaccard Similarity menghitung jumlah dua item yang berurutan (bigram). Algoritma Jaccard Similarity kemudian membagi jumlah bigram yang sama dengan jumlah total bigram unik dalam dua dokumen untuk menghasilkan nilai kesamaan. Algoritma Jaccard Similarity diekspresikan dengan Persamaan dibawah, di mana A dan B adalah dua set string yang diukur berdasarkan tingkat kesamaannya. Nilai Jaccard berkisar dari 0 hingga 1. Algoritma dalam kelas **q-grams**, termasuk Jaccard Similarity, sangat berguna untuk mengatasi kesalahan tipografi tetapi tidak dapat mengukur kesamaan string jika urutannya berubah.

Jaccard Similarity sangat berguna dalam proses penyaringan kandidat. Algoritma ini mengukur tingkat kesamaan antara dua set kata kunci dengan membandingkan keberadaan kata kunci dalam sebuah teks. Dalam konteks ini, Jaccard Similarity efektif untuk mendeteksi kesamaan berdasarkan kata-kata yang sama yang muncul dalam CV kandidat dan deskripsi pekerjaan. Metode ini menghitung rasio antara jumlah elemen irisan (intersection) dengan jumlah elemen gabungan (union) dari kedua set tersebut.

Rumus Jaccard Similarity:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

Di mana A dan B adalah dua set kata dari CV kandidat dan deskripsi HRD.

Implementasi:

Sebelum penerapan Jaccard Similarity, teks diubah menjadi format numerik dengan menggunakan teknik feature engineering seperti TF-IDF. Hal ini memungkinkan sistem untuk mengidentifikasi kata kunci yang relevan dan menghitung kesamaan secara langsung. Metode ini cocok digunakan untuk menangkap kandidat yang memiliki kesesuaian langsung dengan kebutuhan pekerjaan berdasarkan kata kunci tertentu.

- **Cosine Similarity**

  Algoritma *Cosine Similarity* menghitung jarak antara dua objek. Perhitungan dalam metode ini sering menggunakan vektor (Pikies & Ali, 2019). Cosine sudut antara dua vektor yang diproyeksikan pada bidang multidimensi digunakan untuk menghitung kesamaan antara dua dokumen menggunakan metode CS. Kedua vektor tersebut merupakan kumpulan kata (dalam bentuk array) dari dua dokumen yang dibandingkan. Secara matematis, algoritma *Cosine Similarity* dinyatakan dalam Persamaan dibawah di mana a_i dan b_i adalah komponen dari masing-masing vektor. Nilai *Cosine Similarity* berada dalam rentang 0-1. Metode ini lebih cocok untuk data yang memiliki banyak fitur, seperti representasi numerik dari teks menggunakan TF-IDF atau Word2Vec. Dengan mempertimbangkan bobot kata dan hubungan semantik antar kata, Cosine Similarity dapat menangkap kesamaan meskipun kata yang digunakan berbeda tetapi memiliki makna yang sama.

  Rumus Cosine Similarity:

$$Cosine\ Similarity\ =\ Cos(\theta)\ =\ \frac{A.B}{||A||\ x\ ||B||}$$

$$Cos(\theta)\ =\ \frac{\sum_{i=1}^{n} a_i b_i}{\sqrt{\sum_{i=1}^{n} a_i^2}\sqrt{\sum_{i=1}^{n} b_i^2}}$$

  Di mana A dan B adalah vektor teks hasil representasi TF-IDF atau Word2Vec.

  Implementasi:

  Dengan menggunakan teknik feature engineering seperti Word2Vec, hubungan semantik antar kata diwakili dalam bentuk vektor berdimensi rendah. Cosine Similarity bekerja dengan mengukur orientasi (sudut) antara dua vektor, bukan magnitudo, sehingga cocok untuk menangkap makna kontekstual dari teks. Metode ini memberikan fleksibilitas dalam membandingkan CV kandidat dan deskripsi pekerjaan, bahkan jika kata-kata yang digunakan berbeda tetapi serupa secara semantik.

H. Hasil dan Analisis

Kriteria yang diinginkan oleh perusahaan yaitu:

```
"We are looking for a Data Scientist with extensive experience in data processing and analysis using Python. The ideal candidate should also possess skills in web development technologies such as CSS and HTML to support the creation of interactive data visualizations. A strong understanding of machine learning, including the implementation of predictive models, is highly preferred. We seek an individual capable of translating data into meaningful insights and effectively communicating analysis results to cross-functional teams. Additional expertise in database management, big data processing, and familiarity with cloud computing concepts will be a significant advantage."
```

- Cosine Similarity

```
3 kandidat terbaik berdasarkan cosine similarity menggunakan Word2Vec
ID: 21156767, Similarity: 0.9066283106803894
ID: 24610685, Similarity: 0.8954233527183533
ID: 12011623, Similarity: 0.894099771976471
```

Hasil kode tersebut dapat dilihat bahwa kandidat dengan id 21156767, 24610685, dan 12011623 memiliki similarity tertinggi. Pada id 21156767 tercantum bahwa kandidat memiliki pengalaman dengan Python untuk analisis dan pemrosesan data, keahlian dalam machine learning dan implementasi model prediktif, penggunaan SQL untuk manajemen dan pengolahan data, dan pengalaman dengan teknologi pendukung visualisasi data seperti QlikView. Pada CV tersebut terdapat banyak istilah yang relevan, seperti Data Analysis, Business Intelligence, Machine Learning, dan Data Visualization. Kandidat tersebut memiliki proyek yang relevan dengan kategori Data Scientist Proyek prediksi kanker payudara menggunakan Random Forest Classifier, dan proyek klasifikasi gambar menggunakan Java. Serta memiliki pengalaman bekerja selama 6 tahun lebih di industri software.

Dari CV tersebut, dapat diketahui bahwa kandidat dengan ID 21156767 memiliki keahlian yang sangat sesuai dengan deskripsi pekerjaan untuk posisi Data Scientist. Kandidat ini menunjukkan kompetensi dalam pemrosesan dan analisis data menggunakan Python, pemahaman dan implementasi machine learning, serta pengalaman dalam visualisasi data. Selain itu, keahlian dalam pengembangan solusi berbasis SQL dan teknologi web juga turut mendukung relevansi dengan kebutuhan yang dicari oleh HR. Kombinasi kata kunci ini menjadikan skornya paling tinggi karena sangat sesuai dengan persyaratan posisi tersebut.

Dokumen id 24610685 memiliki banyak kata kunci teknis yang relevan dengan deskripsi pekerjaan, seperti SQL, JavaScript, CSS, dan HTML yang mendukung teknologi pengembangan web dan analisis data. Pengalaman sebagai Business Data Analyst dan Sales Engineer menunjukkan kompetensi dalam data mining, database management, dan data analysis, yang sangat sesuai untuk peran Data Scientist. Selain itu, kandidat dapat memvisualisasikan data dengan Tableau, R, dan SAS yang mendukung kemampuan untuk menyajikan data secara interaktif dan informatif.

Pengalaman kandidat dengan id 24610685 dalam data warehousing dan manajemen basis data besar dengan Oracle dan MySQL menunjukkan kemampuan dalam pemrosesan data skala besar. Terbiasa dengan metodologi kerja Agile dan Scrum mencerminkan kemampuan untuk bekerja secara kolaboratif dengan tim lintas fungsi. Soft skills seperti komunikasi yang baik dan analytical thinking mendukung kemampuan untuk menerjemahkan data menjadi wawasan yang bermakna dan menyampaikan hasil analisis dengan efektif kepada pemangku kepentingan.

Dokumen id 12011623 memiliki keahlian yang relevan dengan deskripsi pekerjaan untuk posisi Data Scientist, seperti pemrograman dengan Python, SQL, HTML, dan C#. Pengalamannya dalam data mining, predictive modeling, dan clustering menunjukkan pemahaman mendalam tentang analisis dan pemrosesan data. Selain itu, kemampuan menggunakan alat visualisasi seperti Tableau dan ggplot memperkuat kemampuannya dalam menyajikan data secara interaktif dan informatif, yang sangat dibutuhkan dalam proses pengambilan keputusan berbasis data. Di sisi lain, kandidat ini memiliki keahlian dalam database management dengan Oracle dan SQL Server, serta pengalaman dalam data warehousing yang relevan untuk pemrosesan data skala besar. Pengalaman mengajar dan kemampuan komunikasi yang baik mendukung kebutuhan untuk menerjemahkan data menjadi wawasan bermakna dan berkolaborasi dengan tim lintas fungsi. Kombinasi dari keterampilan teknis dan soft skills ini menjadikannya kandidat yang kuat untuk posisi Data Scientist.

```
3 kandidat terbaik berdasarkan cosine similarity menggunakan TF-IDF:
ID: 12011623.0, Similarity: 0.2011175450799656
ID: 21156767.0, Similarity: 0.17772683762723146
ID: 18067556.0, Similarity: 0.17451318347082773
```

Dokumen id 12011623 dan 21156767 kembali muncul pada cosine similarity menggunakan TF-IDF tetapi memiliki nilai similarity rendah karena beberapa kata kunci pada CV seperti Tableau, Business Intelligence, dan kata yang lebih spesifik mengenai teknologi mungkin tidak sering muncul dalam seluruh dataset sehingga menyebabkan nilai TF-IDF yang lebih rendah, karena sistem TF-IDF lebih fokus pada kata-kata yang lebih umum dan sering ditemukan di banyak dokumen. Selain itu, TF-IDF mengabaikan makna semantik dan hanya memperhatikan frekuensi kata, sehingga meskipun kata kunci yang relevan muncul dalam teks, jika kata-kata tersebut jarang atau tidak berulang dengan frekuensi tinggi, skor TF-IDF bisa lebih rendah dibandingkan dengan metode berbasis Word2Vec karena Word2Vec lebih baik dalam menangkap hubungan kontekstual antar kata.

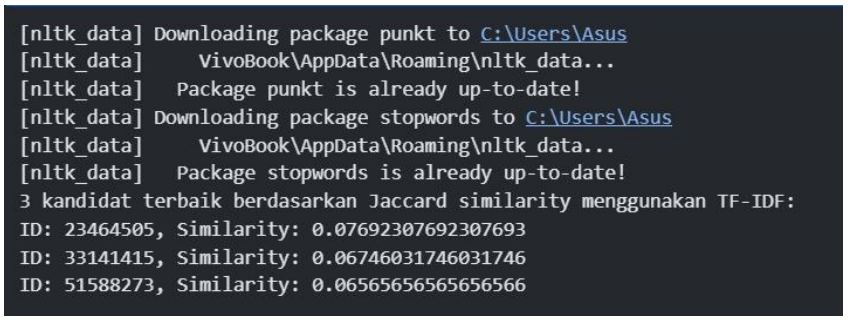
Meskipun Python, Machine Learning, Data Analysis, SQL, dan Business Intelligence tidak sering muncul di seluruh dataset, kata-kata tersebut tetap memiliki frekuensi tinggi dalam dokumen masing-masing. Hal ini membuat kedua CV tersebut tetap relevan dalam perhitungan TF-IDF. Walaupun skor TF-IDF mereka lebih rendah, keduanya masih menunjukkan relevansi yang kuat terhadap deskripsi pekerjaan, dan TF-IDF tetap menilai kata kunci yang sering muncul dalam dokumen tersebut sebagai penting, meskipun dengan nilai similarity yang lebih rendah dibandingkan dengan pendekatan Word2Vec.

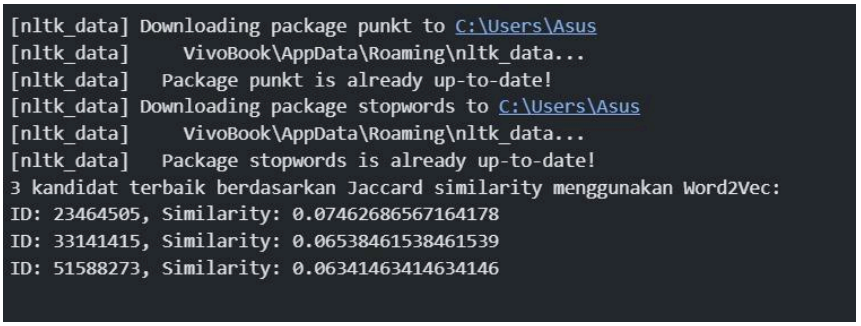
Dokumen ID 18067556 memiliki kesamaan kata kunci dengan deskripsi pekerjaan, seperti Python, data analysis, HTML, CSS, dan database management. Kandidat memiliki pengalaman dalam pemrosesan dan analisis data menggunakan Python serta kemampuan dalam manajemen database dengan SQL, Oracle, dan data warehousing. Selain itu, keterampilan dalam visualisasi data menggunakan PowerBI dan Tableau mendukung kebutuhan untuk menyajikan data secara interaktif. Penguasaan HTML dan CSS disebutkan, meskipun tidak menjadi fokus utama dalam profil kandidat.

Walaupun pemahaman eksplisit tentang machine learning dan implementasi model prediktif tidak disebutkan, pengalaman kandidat dalam analisis bisnis dan proses ETL menunjukkan kemampuan menangani data kompleks dan menerjemahkannya menjadi wawasan yang bermanfaat. Hal ini relevan dengan kriteria untuk komunikasi mengenai hasil analisis ke tim lintas fungsi. Dokumen ID 18067556 muncul karena keterampilan inti dalam pemrosesan data, manajemen basis data, dan kemampuan yang mendukung analisis serta

visualisasi data, meskipun beberapa aspek seperti machine learning dan cloud computing tidak secara langsung ditekankan.

- Jaccard Similarity





Jaccard similarity metode yang mengukur kesamaan antara dua set, dalam proyek ini set adalah kata-kata yang berhubungan antara CV dan deskripsi pekerjaan. Jaccard similarity dengan Word2Vec dan TF-IDF memiliki hasil yang sama karena mendeteksi kata-kata yang mirip, frekuensi yang sama, dan konteks yang serupa, sehingga walaupun Word2Vec dapat menangkap makna kata dan hubungan antar kata hal tersebut tidak dapat mempengaruhi hasil.

Dokumen id 23464505 memiliki banyak kesamaan kata kunci antara pengalaman kandidat dan deskripsi pekerjaan. Kata-kata seperti data analysis, web development, database management, serta teknologi yang digunakan seperti Spring MVC, Hibernate, dan Oracle 11g relevan dengan kriteria yang dicari dalam deskripsi pekerjaan untuk posisi Data Scientist. Meskipun teknologi seperti Python, machine learning, dan predictive models tidak secara eksplisit disebutkan dalam CV, pengalaman kandidat dalam aspek teknis lainnya tetap mendukung kesesuaian dengan posisi yang dimaksud.

Dari hasil eksperimen yang dilakukan menggunakan metode clustering K-Means dan Hierarchical Agglomerative Clustering, dapat disimpulkan bahwa algoritma ini berhasil mengelompokkan kandidat berdasarkan kesamaan keterampilan dan pengalaman yang relevan dengan posisi pekerjaan. Cluster 0 terdiri dari kandidat dengan keahlian di bidang penjualan, pengembangan produk, dan manajemen bisnis, sehingga mereka cocok untuk posisi di bidang penjualan dan pemasaran. Cluster 1, yang berfokus pada manajemen operasional, keuangan, dan sumber daya manusia, lebih relevan untuk posisi manajer keuangan, kepala departemen, atau posisi HR. Sementara itu, Cluster 2 berisi kandidat yang memiliki keterampilan dalam manajemen proyek dan pengelolaan sistem, yang sesuai untuk posisi manajer proyek, pengembang perangkat lunak, atau posisi teknis lainnya.

Metode terbaik yang telah dianalisis adalah penggunaan Cosine Similarity dengan Word2Vec, karena metode ini menghasilkan nilai similarity tertinggi dibandingkan dengan metode lainnya dan paling relevan dengan deskripsi pekerjaan yang dibutuhkan. Metode TF-IDF memberikan hasil yang cukup relevan, metode ini lebih mengandalkan frekuensi kata dan kurang mampu menangkap makna yang lebih dalam di balik kata-kata tersebut, sedangkan Word2Vec mampu mengidentifikasi hubungan makna antara kata-kata yang

mungkin tidak sering muncul dalam teks, seperti dalam kata kunci yang relevan dalam CV dan deskripsi pekerjaan, seperti data analysis, machine learning, dan database management. Oleh karena itu, Word2Vec menghasilkan kecocokan yang lebih akurat dan mencerminkan kesamaan yang lebih mendalam antara CV kandidat dan deskripsi pekerjaan.

I. Kesimpulan dan Saran

Berdasarkan hasil analisis yang telah dilakukan, metode terbaik untuk menyaring dan mengelompokkan kandidat adalah cosine similarity dengan TF-IDF. Metode ini memberikan nilai similarity rendah namun paling relevan dengan syarat pekerjaan yang dibutuhkan. Keunggulan TF-IDF terletak pada kemampuannya menangkap kata yang sama pada sebuah dokumen. Meskipun Word2Vec juga menghasilkan nilai similarity yang tinggi dan memberikan hasil yang cukup relevan, namun metode ini lebih mengandalkan semantik setiap kata dan kurang mampu mencari kata yang sama untuk syarat pekerjaan. Selain itu, metode Jaccard Similarity membantu mengidentifikasi kesamaan berdasarkan keberadaan kata kunci secara langsung, namun kurang efektif dalam menangkap hubungan semantik antar kata.

Eksperimen menunjukkan bahwa kandidat dengan ID 21156767, 12011623 , dan 18067556 memiliki similarity tertinggi pada Cosine Similarity menggunakan TF-IDF dengan deskripsi pekerjaan untuk posisi Data Scientist. Kandidat ini memiliki kombinasi keahlian yang mencakup pemrosesan data dengan Python, machine learning, database management, dan visualisasi data, yang sesuai dengan kebutuhan posisi tersebut. Di sisi lain, hasil dari Jaccard Similarity menunjukkan bahwa kandidat dengan ID 23464505 juga relevan karena memiliki banyak kesamaan kata kunci dengan deskripsi pekerjaan, meskipun tidak secara eksplisit menyebutkan machine learning atau predictive models.

Berdasarkan hasil eksperimen, terdapat kendala pada representasi fitur dan keragaman data CV. Representasi fitur yang menggunakan TF-IDF dan Word2Vec masih memiliki keterbatasan dalam mengukur kesamaan antar kandidat secara menyeluruh. TF-IDF memperhitungkan frekuensi kemunculan kata tanpa mempertimbangkan makna dari kata tersebut sehingga menyebabkan nilai similarity rendah, sedangkan Word2Vec terkadang tidak mampu menangkap konteks secara mendalam karena semantik setiap kata dihitung atau dipertimbangkan. Selain itu, variasi yang signifikan dalam format, istilah, dan fokus keahlian pada CV kandidat menyulitkan proses penyaringan dan pengelompokan. Perbedaan ini mempengaruhi keakuratan dalam mendeteksi kesamaan antara CV dan deskripsi pekerjaan, sehingga mengurangi efektivitas metode yang digunakan untuk menentukan relevansi kandidat.

Referensi

Cahyanti, F. E., Adiwijaya, & Faraby, S. Al. (2020). On the Feature Extraction for Sentiment Analysis of Movie Reviews Based on SVM. *2020 8th International Conference on Information and Communication Technology, ICoICT 2020*. https://doi.org/10.1109/ICOICT49345.2020.9166397

Diana, N. E., & Hanana Ulfa, I. (2019). Measuring performance of n-gram and jaccard-similarity metrics in document plagiarism application. *Journal of Physics: Conference Series*, *1196*(1). https://doi.org/10.1088/1742-6596/1196/1/012069

Gali, N., … R. M.-I.-2016 23rd I., & 2016, undefined. (n.d.). Similarity measures for title matching. *Ieeexplore.Ieee.OrgN Gali, R Mariescu-Istodor, P Fränti2016 23rd International Conference on Pattern Recognition (ICPR), 2016•ieeexplore.Ieee.Org*. Retrieved December 26, 2024, from https://ieeexplore.ieee.org/abstract/document/7899857/

Hasan, T., Matin, A., Kamruzzaman, M., Islam, S., Hasan, T., Matin, A., Kamruzzaman, M., Islam, S., & Goni, M. O. F. (n.d.). A comparative analysis of feature extraction methods for human opinion grouping using several machine learning techniques. *Ieeexplore.Ieee.OrgT Hasan, A Matin, M Kamruzzaman, S Islam, MOF Goni2020 IEEE International Women in Engineering (WIE) Conference on, 2020•ieeexplore.Ieee.Org*. https://doi.org/10.1109/WIECON-ECE52138.2020.9398025

Muhammad, P., … R. K.-P. C., & 2021, undefined. (n.d.). Sentiment analysis using Word2vec and long short-term memory (LSTM) for Indonesian hotel reviews. *ElsevierPF Muhammad, R Kusumaningrum, A WibowoProcedia Computer Science, 2021•Elsevier*. Retrieved December 26, 2024, from https://www.sciencedirect.com/science/article/pii/S1877050921000752

Nawangsari, R., … R. K.-P. C., & 2019, undefined. (n.d.). Word2vec for Indonesian sentiment analysis towards hotel reviews: An evaluation study. *ElsevierRP Nawangsari, R Kusumaningrum, A WibowoProcedia Computer Science, 2019•Elsevier*. Retrieved December 26, 2024, from https://www.sciencedirect.com/science/article/pii/S1877050919310968

Pikies, M., & Ali, J. (2019). String similarity algorithms for a ticket classification system. *2019 6th International Conference on Control, Decision and Information Technologies, CoDIT 2019*, 36–41. https://doi.org/10.1109/CODIT.2019.8820497

Raflika, L., Amanda Putri, R., Yasir Ardiansyah, M., Islam Negeri Sumatera Utara -Medan Jl William Iskandar PsV, U., Estate, M., & Percut Sei Tuan, K. (2024). REKRUTMEN DAN SELEKSI SUMBER DAYA MANUSIA DI ERA DIGITAL. *Musytari : Jurnal Manajemen, Akuntansi, Dan Ekonomi*, *11*(8), 31–40. https://doi.org/10.8734/MUSYTARI.V11I8.8246

Ranjan, S., & Mishra, S. (2020). Comparative Sentiment Analysis of App Reviews. *2020 11th International Conference on Computing, Communication and Networking Technologies, ICCCNT 2020*. https://doi.org/10.1109/ICCCNT49239.2020.9225348

Shi, B., Zhao, J., & Xu, K. (2019). A word2vec model for sentiment analysis of weibo. *2019 16th International Conference on Service Systems and Service Management, ICSSSM 2019*. https://doi.org/10.1109/ICSSSM.2019.8887652

Wahidin, D., Rosmaladewi, O., & Janaenah. (2024). Implementasi Artificial Intelligence dalam Proses Rekrutmen dan Seleksi : Studi Kasus pada PT Saripetejo Sukabumi. *Jurnal Indragiri Penelitian Multidisiplin*, *4*(3), 84–88. https://ejournal.indrainstitute.id/index.php/jipm/article/view/1074

KONTRIBUSI ANGGOTA

No	**Nama**	**Kontribusi**
1	Hasna Lestari (23031554033)	Membuat Laporan Progress dan Laporan Akhir, Mencari Jurnal Terkait
2	Arfan Hafidt Nashrullah (23031554112)	Feature Engineering, Text Similarity (Cosine, Jaccard), Clustering
3	Muhammad Fariz Abid R (23031554030)	Merge Data, Preprocessing, Membuat Presentasi PowerPoint