Prediksi Keberhasilan Lamaran Pekerjaan Dengan *Count Vectorizer* dan *Logistic Regression*

Ellina[1], Jasen Cristian[2], Valerie Louise[3], Sabrilla Koka[4], Christnatalis[5]

Universitas Prima Indonesia

[1]ellinaphan0606@gmail.com, [2]jasenc795@gmail.com, [3]valerielouise18@gmail.com, [4]sabrillakk@gmail.com, [5]chrisnatalis@unprimdn.ac.id

Abstract

Job application letters are commonly submitted in the job application process, therefore the process of assessing job applicants can take a lot of time and effort. The use of Machine Learning technology that relies on learning techniques like humans is a solution that can be done to help the analysis or prediction process. Logistic Regression is one of the Machine Learning algorithms that can classify true and false values, also known as the Binary Classification. In this study, the authors implement the algorithm to predict whether a job application letter will be accepted by the company or not. The dataset used has a total of 2,155 text data that need to be processed in order for the algorithm to process it. Count Vectorizer is a Natural Language Processing method that can process the text data into matrix data. Therefore, the Machine Learning model can finally understand the meaning behind the text. In the end, this study resulted in prediction accuracy of 85.62%.

Keywords: *Logistic Regression; Machine Learning; Binary Classification; Count Vectorizer; Job Application Prediction*

Abstak

Surat lamaran pekerjaan adalah hal yang umum diajukan dalam proses melamar pekerjaan namun proses penilaian pelamar kerja dapat memakan waktu dan tenaga yang tidak sedikit. Pemanfaatan teknologi Machine Learning yang mengandalkan teknik pembelajaran seperti layaknya manusia adalah solusi yang dapat dilakukan untuk membantu proses analisa ataupun prediksi. Logistic Regression merupakan salah satu algoritma Machine Learning yang dapat melakukan klasifikasi biner antara nilai yang benar dan salah. Dalam penelitian ini, penulis mengimplementasikan algoritma tersebut untuk memprediksi apakah sebuah surat lamaran pekerjaan akan diterima oleh perusahaan atau tidak. Dataset yang digunakan memiliki jumlah 2.155 data, Dikarenakan oleh tipe data surat lamaran berupa teks, maka dapat dilakukan proses Count Vectorizer untuk mengubah format data teks menjadi data matriks yang dapat dimengerti oleh algoritma Logistic Regression. Pada akhirnya, penelitian ini menghasilkan akurasi prediksi sebesar 85,62%.

Keywords: Logistic Regression; Machine Learning; Binary Classification; Count Vectorizer; Prediksi Surat Lamaran Pekerjaan

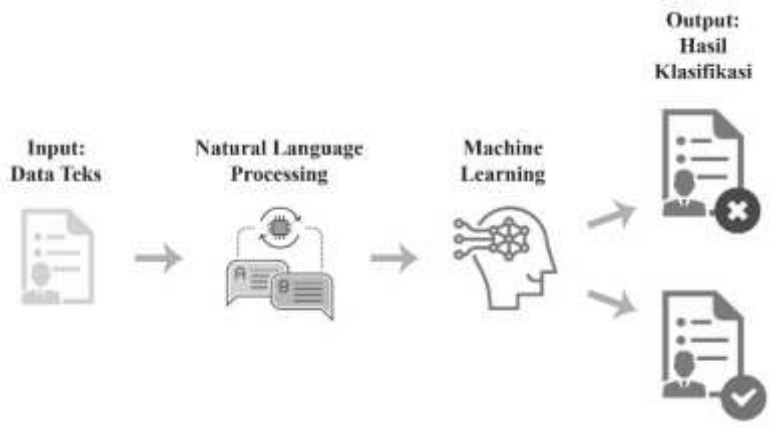
1. Pendahuluan

Metode yang dipakai perusahaan dalam melakukan pemilihan karyawan telah mengalami digitalisasi. Misalnya e-recruitment yang mengubah lamaran kerja dari yang pada awalnya berupa kertas-kertas dalam amplop kini diwakilkan dengan surat digital yang dikirim melalui internet. Masalah yang ditimbulkan apabila perusahaan tidak berhasil mempekerjakan kandidat yang tepat dapat menimbulkan berbagai kerugian seperti kerugian berupa kerugian materi, waktu, penurunan kinerja perusahaan, dan lainnya. Dengan demikian, memasuki era industri 4.0 ini peran teknologi perlu dimanfaatkan pula dalam

ISSN: 2686-0260

proses pemilihan karyawan, salah satunya adalah kecerdasan buatan. Dengan bantuan kecerdasan buatan, proses pemilihan karyawan yang panjang dan memerlukan analisa yang rumit dapat menjadi lebih sederhana dan berkualitas. Selain mendapatkan karyawan yang tepat, perusahaan juga dapat bersaing dan meningkatkan kinerja perusahaan. [1]

Penelitian ini bertujuan untuk memanfaatkan proses Text Classification seperti yang diilustrasikan pada Gambar 1. Text classification pada penelitian ini mengklasifikasikan data surat lamaran pekerjaan yang akan diterima oleh perusahaan atau tidak. Proses klasifikasi dilakukan dengan bantuan algoritma Machine Learning (ML), sedangkan pemrosesan data teks dilakukan dengan bantuan teknik Natural Language Processing (NLP). ML dan NLP termasuk dalam ruang lingkup studi kecerdasan buatan yang dirancang untuk memecahkan masalah yang biasanya hanya dapat diselesaikan oleh manusia. Penelitian sebelumnya juga telah meneliti masalah serupa untuk memprediksi hal tersebut menggunakan teknik NLP yakni Count Vectorizer. Count Vectorizer digunakan untuk memproses data teks lamaran pekerjaan dan algoritma Machine Learning bernama Latent Dirichlet Allocation (LDA) untuk melakukan klasifikasi biner dengan nilai diterima (true) atau tidaknya (false) pelamar kerja tersebut. Hasilnya adalah kemampuan prediksi dengan akurasi sebesar 97% [2]. Teknik prediksi dengan NLP dan Machine Learning sendiri telah dipakai juga untuk memecahkan masalah lainnya seperti mengklasifikasikan artikel [3] dan emosi pada data teks [4].



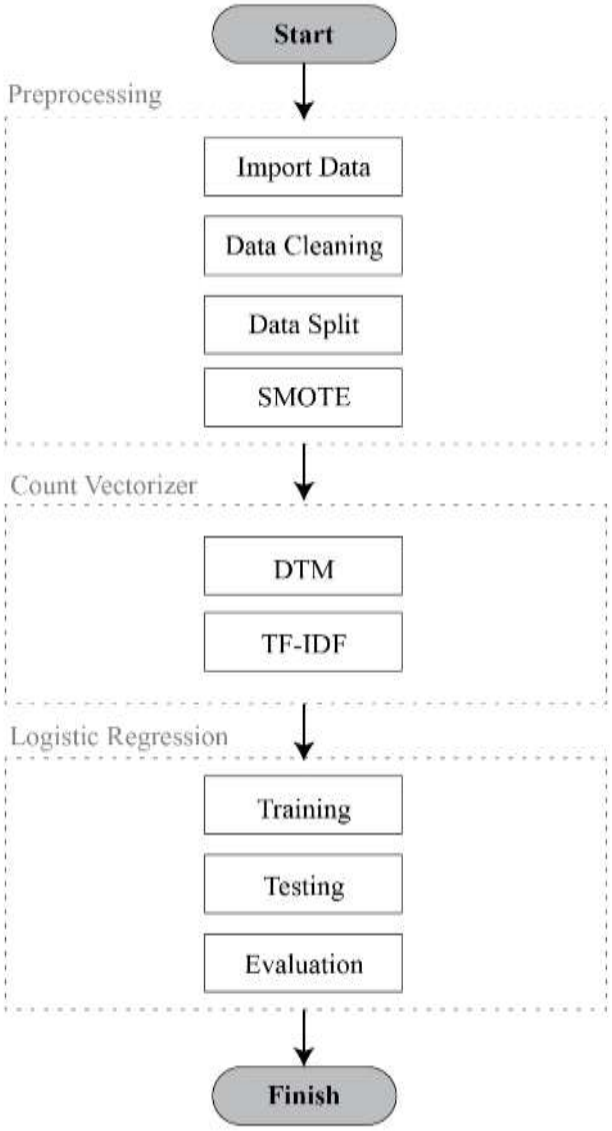
Gambar 1. Proses Text Classification

Selain algoritma LDA, algoritma Machine Learning lainnya telah digunakan untuk memprediksi masalah klasifikasi serupa antara lain Random Forest [5], Logistic Regression [6] dan Multinomial Naïve Bayes [7]. Penelitian ini menggunakan Logistic Regression sebagai algoritma Machine Learning karena algoritma ini dinilai cukup efisien, mudah untuk digunakan dan diinterpretasikan. Namun, tidak banyak penelitian yang menggunakan algoritma tersebut terhadap masalah serupa. Oleh karena itu penelitian ini akan menguji kemampuan algoritma tersebut. Kemudian, selain Count Vectorizer, juga terdapat metode NLP lainnya seperti Word2Vec yang menggunakan metode Neural Network untuk mempelajari data teks dan Node2Vec yang menggunakan representasi vektor dari nodes pada sebuah grafik [3]. Penelitian ini akan menggunakan teknik Count Vectorizer karena teknik ini dinilai cukup baik dalam memproses data teks menjadi data matriks yang dapat diproses oleh algoritma Machine Learning.

Jumlah data yang diteliti pada penelitian ini adalah sebanyak 2.155 data yang didapatkan pada tahun 2020 hingga 2021. Sedangkan jenis klasifikasi yang dilakukan pada penelitian ini adalah binary classification atau klasifikasi dengan dua output. Kedua output yang diteliti pada penelitian ini adalah berhasil atau tidaknya seorang pelamar kerja yang disimbolkan dengan variabel dependen accepted dengan nilai 1 (true) dan 0 (false). Selain itu, data lamaran pekerjaan juga telah diteliti untuk meneliti keberhasilan pelamar kerja [8],

dan juga untuk memprediksi jenis karir [9]. Pada penelitian ini data yang diteliti didapatkan dari sebuah situs layanan online untuk melamar pekerjaan. Secara garis besar, data tersebut berupa teks atau kata-kata yang kemudian diproses dengan metode Count Vectorizer dan diprediksi dengan algoritma Machine Learning Logistic Regression untuk memprediksi diterima atau tidaknya seorang pelamar kerja.

2. Metodologi Penelitian



Gambar 2. Flowchart Proses Prediksi Surat Lamaran Pekerjaan

Penelitian ini menggunakan metode pengumpulan data dengan mengumpulkan data surat lamaran pekerjaan yang hendak diteliti, serta analisis eksperimental menggunakan program Jupyter Notebook dan bahasa pemrograman Python 3 untuk merancang model Machine Learning yang memprediksi keberhasilan surat lamaran pekerjaan dengan algoritma Logistic Regression. Adapun beberapa library yang digunakan antara lain Pandas untuk mengolah dataset, Sci-kit Learn untuk menggunakan algoritma Logistic Regression, dan NLTK untuk melakukan proses Count Vectorizer. Selanjutnya secara garis besar, terdapat tiga prosedur utama dalam penelitian ini, yakni antara lain 1) Data Pre-processing, 2) Count Vectorizer, dan 3) Logistic Regression seperti yang diilustrasikan pada Gambar 1.

2.1. Preprocessing

Preprocessing adalah proses pemrosesan data sebelum dapat digunakan oleh algoritma seperti Count Vectorizer dan Logistic Regression. Proses ini juga penting karena dapat mempengaruhi performa akurasi dalam memprediksi surat lamaran pekerjaan.

1. Import Data

Tahap pertama pada proses preprocessing dimulai dengan mengimpor data surat lamaran kerja yang memiliki format .CSV (Comma Separated Value) ke dalam Aplikasi Jupyter Notebook dengan format dataframe. Hal ini diperlukan agar data dapat dibaca oleh bahasa pemrograman Python dan algoritma pemrograman.

2. Data Cleaning

Tahap selanjutnya adalah Data Cleaning yang bertujuan untuk membersihkan data agar dapat diproses ke tahap selanjutnya:

a. Penghapusan Stopwords
   Stopwords adalah kata-kata yang tidak menambah makna keseluruhan dari teks kita, contohnya adalah kata sambung seperti 'pada', 'adalah', dan 'antara lain'. Kata-kata tersebut akan dihapus untuk memastikan data yang digunakan adalah data yang memiliki makna.
b. Penghapusan Punctuation
   Punctuation atau tanda baca menambahkan noise atau kesalahan pemahaman makna yang membawa ambiguitas saat melatih model. Hal ini dikarenakan tanda baca dapat disalah artikan sebagai suatu karakter yang memiliki makna, padahal sebenarnya tidak. Oleh karena itu penting untuk menghapus tanda baca pada data yang hendak diproses.
c. Penghapusan data duplikasi
   Penghapusan data duplikasi penting karena dapat mengakibatkan algoritma Logistic Regression kebingungan dalam menganalisa data surat lamaran pekerjaan yang sama berulang-ulang.
d. Penghapusan data dengan jumlah kata dibawah 10
   Data dapat mengandung beberapa data yang memiliki jumlah kata yang terlalu sedikit dan dapat menghasilkan penurunan akurasi saat memprediksi data tersebut sehingga data tersebut perlu dibersihkan.

3. Data Split

Proses ini membagi dataset menjadi dua bagian, yakni Testing Data sebesar 25% dan Training Data sebesar 75%. Proses ini penting untuk memastikan data yang hendak dipakai dalam proses Testing atau pengujian berbeda dengan data yang terdapat pada proses pelatihan atau Training. Masalah yang timbul apabila proses ini tidak dilakukan adalah algoritma akan handal dalam menganalisa data Training namun tidak dapat menganalisa data baru yang belum pernah dipelajari.

4. SMOTE

Synthetic Minority Over-sampling Technique (SMOTE) merupakan metode yang dapat digunakan untuk menangani ketidak seimbangan data terutama saat digunakan untuk permasalahan Data Science. Teknik ini melakukan sintesis sampel dari data minoritas untuk menyeimbangkan dataset melalui proses pengambilan sampel dari kelas minoritas [10].

2.2. Count Vectorizer

Setelah data berhasil dibersihkan dan menghasilkan Training Data dan Testing Data, maka selanjutnya proses Count Vectorizer dilakukan untuk mentransformasi data teks surat lamaran pekerjaan menjadi data vektor berbasis frekuensi dari setiap kata yang muncul di

dalam teks tersebut. Hal ini penting karena algoritma Logistic Regression tidak dapat memanfaatkan teks tersebut dengan baik sebagai variabel independen tanpa melalui proses ini. Terdapat dua tahap dalam proses Count Vectorizer pada penelitian ini, antara lain:

1. DTM

Setelah proses Count Vectorizer selesai maka dihasilkan output berupa data dengan tipe data Document Term Matrix (DTM). DTM adalah matriks matematis yang menggambarkan frekuensi kemunculan istilah dalam kumpulan dokumen. Dalam DTM, baris merepresentasikan dokumen teks dan kolom merepresentasikan istilah pada teks tersebut.

2. TF-IDF

Setelah proses Document Term Matrix dilakukan, maka selanjutnya dapat dilakukan proses Term-Frequency Times Inverse Document-Frequency (TF-IDF). TF-IDF adalah ukuran statistik yang mengevaluasi seberapa relevan sebuah kata dengan dokumen dalam kumpulan dokumen. Hal ini dilakukan dengan mengalikan dua metrik: 1) Berapa kali sebuah kata muncul dalam sebuah dokumen, dan 2) Frekuensi dokumen terbalik dari kata tersebut di seluruh kumpulan dokumen atau teks tersebut.

Dengan demikian, dapat dicontohkan apabila terdapat dua teks berikut:

1. "Saya suka olahraga"
2. "Saya juga suka olahraga "

maka contoh output yang dihasilkan adalah seperti yang dicontohkan pada Tabel 1.

Tabel 1. Contoh DTM & TF-IDF

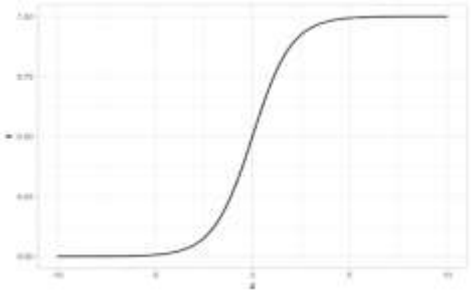
	Saya	**Juga**	**Suka**	**olahraga**
DTM-1: Saya suka olahraga	1	0	1	1
DTM-2: Saya juga suka olahraga	1	1	1	1
TF-IDF-1: Saya suka olahraga	0.704	0.000	0.707	0.707
TF-IDF-2: Saya juga suka olahraga	0.704	0.501	0.501	0.501

2.3. Logistic Regression

Logistic regression adalah bentuk khusus regresi yang diformulasikan untuk melakukan klasifikasi data dependen yang memiliki dua group prediksi (misalnya true dan false). Dimana Variabel independen yang dilambangkan dengan Y adalah variabel yang nilainya tergantung dari variabel independen yang nilainya dapat berubah. Di lain sisi, variabel independen adalah variabel yang menyebabkan atau mempengaruhi perubahan variabel dependen. Variabel ini dilambangkan dengan X_1, X_2, dan seterusnya. Dengan diketahuinya kedua variabel tersebut maka bentuk umum variabel dependen dan independen pada logistic regression dapat diilustrasikan sebagai berikut:

$$Y = X_1 + X_2 + X_3 + \ldots + X_n \quad (1)$$

Fungsi logistik standar (atau fungsi sigmoid) seperti yang diilustrasikan pada Gambar 2.2 merupakan fungsi yang dipakai oleh Logistic Regression.



Gambar 3. Ilustrasi Fungsi Sigmoid

Misalkan z adalah sembarang nilai kontinu yang domainnya adalah (-∞, ∞). Jika kita memasukkan z ke fungsi sigmoid maka akan seperti rumus berikut:

$$\theta(z) = \frac{e^z}{1+e^z} = \frac{1}{1+e^{-z}} \quad (2)$$

Model Evaluation bertujuan untuk mengetahui performa akurasi Machine Learning dalam memprediksi surat lamaran pekerjaan. Dengan memberikan label True dan False terhadap nilai prediksi yang sesuai dengan fakta (True) dan yang tidak sesuai dengan fakta (False). Kemudian memberikan label Positive dan Negative terhadap surat lamaran pekerjaan yang diterima oleh perusahaan (Positive) dan tidak diterima oleh perusahaan (Negative). Maka didapatkan 4 kuadran yang disebut dengan Confusion Matrix. Confusion Matrix terdiri:

TP = jumlah lamaran pekerjaan yang diterima dan sesuai antara fakta dengan prediksi

TN = jumlah lamaran pekerjaan yang ditolak dan sesuai antara fakta dengan prediksi

FP = jumlah lamaran pekerjaan yang diterima dan tidak sesuai antara fakta dengan prediksi

FN = jumlah lamaran pekerjaan yang ditolak dan tidak sesuai antara fakta dengan prediksi

Dengan demikian rumus untuk menghitung akurasi prediksi adalah:

$$\text{Akurasi} = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \quad (3)$$

3. Hasil Dan Pembahasan

3.1. Hasil Penelitian

Persiapan awal dalam penelitian ini adalah dengan mengimpor 2.155 data yang akan diteliti dan library-library yang dibutuhkan. Keterangan mengenai data pada penelitian ini ditampilkan pada Tabel 2.

Tabel 2. Keterangan Data Sebelum Preprocessing

Jumlah Data	2.155 data
Variabel Dependen	Accepted (Diterima atau tidaknya sebuah surat lamaran pekerjaan)
Jumlah Accepted bernilai True	304 data
Jumlah Accepted bernilai False	1.851 data

Variabel Independen	Cover Letter (Surat lamaran pekerjaan)
Bahasa yang digunakan	Bahasa Inggris
Panjang rata-rata variabel independen	329 kata
Panjang minimal variabel independen	10 kata
Panjang maksimal variabel independen	556 kata

Mula-mula proses Preprocessing dilakukan untuk membersihkan data surat lamaran pekerjaan dari Import Data, Data Cleaning, Data Split, hingga proses SMOTE. Setelah data dibersihkan maka bentuk dataset menjadi seperti yang dideskripsikan pada Tabel 3.

Tabel 3. Keterangan Data Setelah Preprocessing dan Count Vectorizer

Jumlah Data	3.086 data
Jumlah Accepted bernilai True	1.543 data
Jumlah Accepted bernilai False	1.543 data
Jumlah Data Train	2.314 data
Jumlah Data Test	772 data
Panjang rata-rata variabel independen	208 kata
Panjang minimal variabel independen	9 kata
Panjang maksimal variabel independen	445 kata

Kemudian selanjutnya, hasil Training dan Testing untuk memprediksi keberhasilan surat lamaran pekerjaan ditampilkan pada Tabel 4. Variabel dependen yakni 'Accepted' memiliki dua kemungkinan, yaitu 0 (false) dan 1(true). Nilai 0 menandakan surat lamaran yang tidak diterima perusahaan dan 1 merupakan surat lamaran yang diterima oleh perusahaan. Terlihat pada tabel hasil tersebut bahwa terdapat kesalahan prediksi pada index 28 dengan nilai prediksi 0 (tidak diterima perusahaan) padahal seharusnya 1 (diterima oleh perusahaan). Kemudian sebaliknya, hasil prediksi yang tepat dapat dicontohkan di baris lainnya dimana nilai prediksi dan aktual adalah sama.

Tabel 4. Hasil Prediksi Surat Lamaran Pekerjaan

Index	Variabel Independen Setelah Preprocessing	Nilai Prediksi[a]	Nilai Aktual[a]
419	done digital marketing previous internship kno...	0	0
2184	Well driven creative eager learn technological...	0	0
606	young professional willing learn teachable ind...	1	1
1677	effortless determined person always willing le...	0	0
1761	think suitable background marketing love learn…	0	0
…	…	…	
2133	leaner chllange taker job alomst new role lear...	0	0
2022	rich practical experience worked Internet indu...	0	0
765	believe suitable job love taking quality photo...	0	0

28	Dear super clean team writing express interest...	0 [b]	1
1660	experience shift supervisor well field digital...	0	0

a: 1=Diterima oleh perusahaan, 0=Tidak diterima perusahaan

b: Hasil prediksi yang salah

Setelah proses Training dan Testing selesai maka proses terakhir yakni Evaluation, dilakukan menggunakan bantuan library Sci-kit Learn accuracy_score dan confusion_matrix seperti pada Gambar 4.

```
cm = confusion_matrix(y_test_res, y_pred_class)
cm = pd.DataFrame(
    cm,
    columns = ['Hasil Prediksi False','Hasil Prediksi True'],
    index = ['Hasil Aktual False','Hasil Aktual True']
)
print('Confusion matrix')
print(cm)

print(f'\nAkurasi {metrics.accuracy_score(y_test_res, y_pred_class) * 100} %')
```

```
Confusion matrix
                  Hasil Prediksi False  Hasil Prediksi True
Hasil Aktual False                   323                   63
Hasil Aktual True                     48                  338

Akurasi 85.62176165803109 %
```

Gambar 4. Hasil Evaluasi

3.2. Analisa Masalah

Penelitian ini berhasil membuktikan bahwa menilai surat lamaran pekerjaan menggunakan teknologi adalah hal yang dapat dilakukan. Tantangan yang muncul antara lain:

1. Masalah sulitnya menggunakan tipe data berupa teks atau paragraf untuk diprediksi oleh algoritma yang hanya bisa membaca data berupa angka. Selain itu, algoritma pun perlu mempelajari tata bahasa pada surat lamaran pekerjaan. Penelitian ini membuktikan bahwa masalah ini dapat diatasi dengan mengimplementasikan teknik Count Vectorizer yang bukan hanya mengubah data teks menjadi angka, namun juga frekuensi seberapa banyak istilah dalam teks muncul.
2. Selain itu masalah ketidakseimbangan data juga merupakan tantangan yang dapat diatasi dengan proses Synthetic Minority Over-sampling Technique (SMOTE).
3. Kemudian yang terakhir, proses penilaian surat lamaran pekerjaan dapat dilakukan dengan teknik pembelajaran dan prediksi yang dimiliki oleh algoritma Machine Learning Logistic Regression.

3.3. Akurasi Logistic Regression

Dengan melihat hasil tabel kebenaran dengan metode Confusion Matrix pada Gambar 3.6 maka dapat diketahui nilai TP = 338, TN = 323, FP = 63, dan FN = 48. Maka tingkat akurasi dapat dihitung sebagai berikut:

$$\text{Akurasi} = \frac{TP + TN}{TP + TN + FP + FN} = \frac{338 + 323}{338 + 323 + 63 + 48} \times 100\% = 85{,}62\%$$

Dengan demikian penelitian ini berhasil memanfaatkan data teks surat lamaran pekerjaan untuk memprediksi keberhasilan lamaran pekerjaan. Beberapa hal yang mendukung algoritma Logistic Regression dapat memprediksi dengan akurasi yang cukup baik sebesar 85,62%, antara lain:

1. Jumlah data yang cukup banyak, yakni sebesar 2.155 data sehingga dapat melatih algoritma dalam memprediksi data yang belum diketahui sebelumnya sekalipun saat memprediksi Testing Data
2. Proses pengolahan data yang berhasil mengolah data teks menjadi data yang dapat diproses oleh algoritma Logistic Regression. Proses ini meliputi Count Vectorization yang mentransformasi data menjadi Document Term Matrix (DTM) dan Term-Frequency Times Inverse Document-Frequency (TF-IDF) yang bukan hanya menghitung jumlah suatu istilah muncul dalam teks, tetapi juga menyusun kosa kata secara otomatis. Selain itu proses SMOTE yang mengatasi masalah Imbalanced Data. Dan yang terakhir pembersihan data berupa penghapusan stopwords untuk menghilangkan kata-kata yang tidak bermakna, penghapusan punctuation untuk mencegah algoritma menyalah artikan tanda baca sebagai suatu istilah, dan penghapusan duplikasi data untuk menghapus data surat lamaran pekerjaan yang sama.
3. Kemampuan logistic regression sebagai algoritma Machine Learning yang dapat memprediksi dengan baik juga turut berperan dalam penelitian ini.

3.4. Perbandingan Hasil Temuan

Bagian ini akan membahas mengenai kaitan antara hasil penelitian ini dengan penelitian sebelumnya. Kecerdasan buatan telah diteliti di penelitian sebelumnya sebagai solusi untuk meningkatkan efisiensi perusahaan dan penelitian ini membuktikan bahwa hal tersebut memang dapat dilakukan [1], [11]. Penelitian ini memanfaatkan teknik kecerdasan buatan yakni Natural Language Processing (NLP) dan Machine Learning (ML) yang terbukti dapat melakukan proses pemilihan karyawan hanya dengan input berupa teks surat lamaran pekerjaan. Hal-hal yang mendukung keberhasilan penelitian ini antara lain proses SMOTE, Logistic Regression, dan Count Vectorizer.

Proses SMOTE pada penelitian ini berhasil mengatasi ketidak-seimbangan data tanpa menurunkan tingkat akurasi secara drastis. Hal tersebut sesuai dengan penelitian sebelumnya [10] yang juga berhasil menggunakan SMOTE untuk mengatasi masalah ketidak-seimbangan data. Kemudian selanjutnya, algoritma NLP yang digunakan pada penelitian ini yakni Count Vectorizer, juga berhasil memproses data teks menjadi data matrix. Hal tersebut juga memperkuat penelitian sebelumnya yang telah berhasil menggunakan algoritma Count Vectorizer untuk memproses data teks [4]. Kemudian yang terakhir, hasil implementasi algoritma Logistic Regression pada penelitian ini juga memperkuat hasil penelitian sebelumnya [11] yang berhasil memprediksi data teks menggunakan algoritma tersebut meskipun akurasi yang dihasilkan tidak terlalu sempurna.

Terdapat beberapa hal yang dapat dikembangkan pada penelitian ini. Meskipun penelitian ini berhasil menghasilkan algoritma Logsitic Regression yang dapat memprediksi keberhasilan surat lamaran pekerjaan, akurasi yang dihasilkan hanyalah 85,62%. Akurasi tersebut tidak sebaik penelitian sebelumnya menggunakan algoritma Latent Dirichlet Allocation (LDA) dengan hasil akurasi yang hampir sempurna sebesar 97% [4]. Hal yang membedakan metode penelitian ini dengan penelitian tersebut adalah kemampuan algoritma LDA yang dibuat khusus untuk menganalisa data teks. Hal tersebut berbeda dengan algoritma Logistic Regression yang merupakan algoritma Machine Learning yang dibuat untuk memprediksi masalah prediktif yang umum dengan kurva logistik.

4. Kesimpulan

Penelitian ini menghasilkan bukti empiris bahwa teknologi Machine Learning dapat memprediksi keberhasilan surat lamaran pekerjaan yang biasanya dilakukan oleh manusia secara manual. Penelitian ini berhasil melakukan prediksi teks surat lamaran pekerjaan yang telah dibersihkan melalui proses Preprocessing dan diproses dengan teknik Natural

Language Processing yakni Count Vectorizer. Namun, terdapat masalah ketidakseimbangan data pada dataset, sehingga proses Synthetic Minority Over-sampling Technique (SMOTE) digunakan untuk menyeimbangkan data. Setelah data telah siap untuk diproses, maka algoritma Logistic Regression dapat digunakan untuk memprediksi variabel dependen berupa diterima atau tidaknya surat lamaran tersebut oleh perusahaan. Hasil prediksi data pengujian (Test) menghasilkan akurasi sebesar 85,62%.

Daftar Pustaka

[1] D. Paramita, "Digitalization in talent acquisition: A case study of AI in recruitment." 2020.

[2] M. Alghazal, "Talent Acquisition Process Optimization Using Machine Learning in Resumes' Ranking and Matching to Job Descriptions," 2021.

[3] M. Grohe, "word2vec, node2vec, graph2vec, x2vec: Towards a theory of vector embeddings of structured data," in *Proceedings of the 39th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*, 2020, pp. 1–16.

[4] D. E. Cahyani and I. Patasik, "Performance comparison of TF-IDF and Word2Vec models for emotion text classification," *Bulletin of Electrical Engineering and Informatics*, vol. 10, no. 5, pp. 2780–2788, 2021.

[5] M. Jayaratne and B. Jayatilleke, "Predicting personality using answers to open-ended interview questions," *IEEE Access*, vol. 8, pp. 115345–115355, 2020.

[6] X. Zhang, J. Kim, R. E. Patzer, S. R. Pitts, A. Patzer, and J. D. Schrager, "Prediction of emergency department hospital admission based on natural language processing and neural networks," *Methods Inf Med*, vol. 56, no. 05, pp. 377–389, 2017.

[7] P. Senarathne, M. Silva, A. Methmini, D. Kavinda, and S. Thelijjagoda, "Automate Traditional Interviewing Process Using Natural Language Processing and Machine Learning," in *2021 6th International Conference for Convergence in Technology (I2CT)*, 2021, pp. 1–6.

[8] F. García-Peñalvo, J. Cruz-Benito, M. Martín-González, A. Vázquez-Ingelmo, J. C. Sánchez-Prieto, and R. Therón, "Proposing a Machine Learning Approach to Analyze and Predict Employment and its Factors," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 5, no. 2, p. 39, 2018, doi: 10.9781/ijimai.2018.02.002.

[9] M. Sippy, J. Khandelwal, A. Jain, and K. K. Mathew, "ResumeScan: Application Tracking and Career Prediction Model," 2021.

[10] R. Siringoringo, "Klasifikasi data tidak seimbang menggunakan algoritma SMOTE dan k-nearest neighbor," *Journal Information System Development (ISD)*, vol. 3, no. 1, 2018.

[11] P. Senarathne, M. Silva, A. Methmini, D. Kavinda, and S. Thelijjagoda, "Automate Traditional Interviewing Process Using Natural Language Processing and Machine Learning," in *2021 6th International Conference for Convergence in Technology (I2CT)*, 2021, pp. 1–6.