Journal of Technology and Informatics (JoTI)
Vol. 7, No. 1, Bulan April Tahun 2025
P-ISSN 2721-4842
E-ISSN 2686-6102

Pemanfaatan Machine Learning dalam Prediksi Rating: Studi Kasus pada Data Abstrak Publikasi Ilmiah

Rizky Zaqi Megantara[1], Pace Iryanto Faot[2], Ridolof Haba Ito[3], Kathleen Felicia Annabel[4], Oscar Karnalim[5*]
[1,2,3,4,5]Program Studi Magister Ilmu Komputer, Fakultas Teknologi dan Rekayasa Cerdas, Universitas Kristen Maranatha, Bandung, Jawa Barat, Indonesia,
e-mail: 2479002@maranatha.ac.id[1], 2479005@maranatha.ac.id[2], 2479004@maranatha.ac.id[3], 2379010@maranatha.ac.id[4], oscar.karnalim@it.maranatha.edu[5*]

Informasi Artikel

***Article History*:**
Received : 17 Februari 2025
Revised : 23 Maret 2025
Accepted : 27 April 2025
Published : 29 April 2025

***Korespondensi:**
oscar.karnalim@it.maranatha.edu

Keywords:
K-Nearest Neighbors (KNN), *Machine Learning*, *Random Forest, Rating Prediction, Support Vector Regressor* (SVR), *XGBoost.*

10.37802/joti.v7i1.999

Journal of Technology and Informatics (JoTI)
P-ISSN 2721-4842
E-ISSN 2686-6102
https://e-journals.dinamika.ac.id/index.php/joti

***Abstract*:**

As the volume of scientific publications increases, the need for automated approaches to evaluate and analyze abstracts becomes increasingly important. This research not only aims to predict the abstract rating of scientific publications using machine learning algorithms, but also offers a unique approach by integrating regression and classification analysis to evaluate the relevance of abstracts more comprehensively. Four main models, namely XGBoost Regressor, Random Forest Regressor, Support Vector Regressor (SVR), and K-Nearest Neighbors Regressor (KNN), are evaluated for this task. The dataset is processed through preprocessing stages which include removing duplications, text representation using TF-IDF, handling data imbalances with Synthetic Minority Oversampling Technique (SMOTE), and dimension reduction using Truncated Singular Value Decomposition (SVD). The research results show that SVR provides performance the best with the lowest Mean Absolute Error (MAE) value of 0.4980, Mean Squared Error (MSE) of 0.5237, and the highest R^2 of 0.7321. XGBoost and Random Forest show competitive performance with advantages in computational efficiency and prediction stability respectively, while KNN provides varying results depending on the data distribution. Dimensionality reduction using Truncated SVD successfully preserves more than 70% of the initial variance, enabling higher computational efficiency without losing important information. This research makes a significant contribution in supporting machine learning-based decision making, especially in the analysis of abstracts of scientific publications. This approach can be further developed through exploration of ensemble or hybrid models, as well as testing on larger datasets to improve generalization and accuracy.

PENDAHULUAN

Kemajuan teknologi informasi telah membawa perubahan signifikan dalam cara peneliti mengakses dan mengevaluasi literatur ilmiah. Informasi terkait artikel ilmiah tidak hanya berada di perpustakaan, tetapi dapat pula ditemukan secara *online*. Salah satu *platform* untuk mencari informasi artikel ilmiah di internet adalah dengan menggunakan *Google Scholar*.

Google Scholar adalah situs pencarian yang secara spesifik menjadi tempat untuk pencarian literatur dan sumber akademik [1] *Google Scholar* menyediakan akses yang luas ke berbagai artikel ilmiah, memungkinkan para peneliti untuk menemukan karya-karya yang relevan dengan bidang studi mereka [2]. Salah satu bagian penting dalam hasil pencarian adalah abstrak yang muncul terkait topik yang dipilih. Abstrak merupakan bagian penting dari artikel ilmiah yang memiliki fungsi untuk memberikan gambaran singkat tentang tujuan, metodologi, dan hasil penelitian. Penilaian terhadap abstrak sering kali menjadi langkah awal bagi pembaca untuk memutuskan apakah artikel tersebut relevan dengan kebutuhan mereka [3].

Namun, meskipun *Google Scholar* menawarkan akses luas ke literatur, proses penilaian relevansi abstrak saat ini masih bergantung pada evaluasi subjektif manusia. Evaluasi ini sering kali memakan waktu, rentan terhadap bias [4], dan menjadi tantangan ketika jumlah artikel yang tersedia sangat besar. Selain itu, sebagian besar sistem pencarian literatur belum menyediakan cara otomatis untuk memberikan penilaian atau rating terhadap relevansi abstrak berdasarkan kriteria tekstual tertentu. *Gap* ini menunjukkan perlunya pendekatan yang lebih otomatis dan objektif untuk mendukung efisiensi dalam seleksi literatur ilmiah.

Studi yang [5] memperkenalkan pendekatan *Modularized Hierarchical Convolutional Neural Network (MHCNN)* untuk mengevaluasi akademik secara otomatis. Studi ini menunjukkan potensi arsitektur *deep learning* untuk memprediksi penerimaan atau penolakan makalah berdasarkan elemen-elemen struktural seperti judul, abstrak, dan kesimpulan. Namun, penelitian tersebut terutama berfokus pada klasifikasi biner *(accept/reject)* dan menggunakan dataset yang mencakup seluruh bagian makalah. Dalam hal ini, studi yang akan dilakukan pada penelitian ini memiliki perbedaan signifikan dengan memberikan kontribusi pada analisis kuantitatif prediksi rating berbasis abstrak, yang mencakup integrasi regresi dan klasifikasi untuk mengevaluasi tingkat relevansi.

Dalam konteks ini, pengembangan model prediktif untuk mengevaluasi relevansi dan memberikan *rating* terhadap abstrak menjadi solusi yang menjanjikan. Dengan memanfaatkan *Natural Language Processing (NLP)* dan algoritma pembelajaran mesin, model ini dapat membantu menyaring dan memprioritaskan literatur ilmiah secara efisien. Sebagai contoh, penelitian [5] menggunakan pendekatan modular untuk menilai makalah akademik, tetapi penelitian ini tidak mencakup analisis mendalam terhadap prediksi numerik rating atau evaluasi relevansi dalam kategori yang lebih spesifik. Oleh karena itu, penelitian ini menawarkan pendekatan yang unik dengan menggunakan regresi untuk memprediksi rating abstrak dan mengklasifikasikan prediksi tersebut menjadi kategori relevan, netral, dan tidak relevan.

Penelitian ini bertujuan untuk membangun model prediksi rating berdasarkan abstrak dari artikel yang diambil dari *Google Scholar*. Model ini akan membandingkan empat macam algoritma, yaitu *XGBoost Regressor, Random Forest Regressor (RF), Support Vector Regressor (SVR) dan K-Nearest Neighbors Regressor (KNNRegressor)* untuk mengevaluasi abstrak berdasarkan fitur-fitur tekstual yang diekstraksi, dengan metrik evaluasi menggunakan *Mean Square Error*, *Mean Average Error, Coefficient of Determination* (R^2), dan *Accuracy*. Hasil penelitian ini diharapkan dapat memberikan kontribusi dalam mempermudah proses evaluasi literatur ilmiah, meningkatkan efisiensi waktu, dan mengurangi potensi bias dalam penilaian abstrak.

Studi Literatur

Berbagai penelitian telah dilakukan dalam analisis data berbasis teks. Pendekatan seperti *XGBoost* dan *Ridge Regression* digunakan untuk memprediksi popularitas *review online*

dengan hasil yang menunjukkan bahwa fitur tekstual memiliki kontribusi lebih signifikan dibandingkan fitur non-tekstual [6]. Dalam analisis sentimen, beragam metode seperti *Latent Semantic Indexing* [7], *ClimateBERT* [8] dan model *hybrid* berbasis *deep learning* [9] menunjukkan peningkatan akurasi dalam memahami pola sentimen, baik untuk isu sosial, bisnis maupun lingkungan. Selain itu, model prediktif berbasis *ensemble* telah berhasil meningkatkan klasifikasi sentimen dalam bahasa spesifik seperti Persia [9]. Penelitian lain juga menunjukkan bahwa integrasi analisis sentimen dengan peramalan berbasis deret waktu dapat meningkatkan akurasi prediksi penjualan hingga 96% [10].

Studi lain mengeksplorasi emosi publik terhadap isu lingkungan selama satu dekade [11], deteksi ujaran kebencian di Twitter [12] dan sentimen selama pandemi COVID-19 [13]. Model BB-DAM untuk analisis lintas domain menunjukkan performa yang unggul [14] dan [15] mengevaluasi respons emosional siswa terhadap pembelajaran ilmu komputer. Chen et al. [16] mengembangkan model prediksi peringkat produk berbasis teks ulasan menggunakan pendekatan *bag-of-words* dengan *unigram* yang terbukti lebih akurat daripada *bigram*, meningkatkan kinerja prediksi hingga 15,89%. Pendekatan ini menunjukkan potensi analisis berbasis teks sebagai alat utama dalam sistem rekomendasi.

Pendekatan serupa diterapkan oleh Isah et al. [17] untuk menganalisis keamanan produk melalui media sosial. Dengan menggunakan metode berbasis leksikon dan pembelajaran mesin seperti *Naive Bayes* dan SVM, penelitian ini membuktikan efektivitas pengolahan data media sosial dalam menilai sentimen terhadap merek dan produk. Mukherjee et al. [18] melangkah lebih jauh dengan analisis fitur spesifik menggunakan *dependency parsing* untuk mengidentifikasi hubungan antara opini dan fitur produk, menghasilkan akurasi yang tinggi pada berbagai dataset.

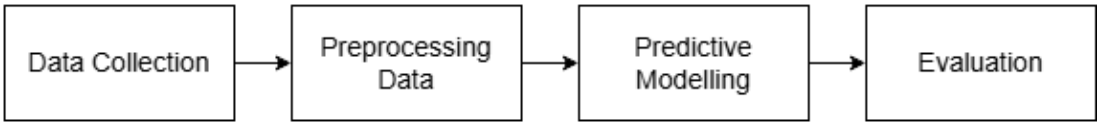
Visualisasi sentimen menjadi fokus penelitian Bhatt et al. [19], yang menyederhanakan ulasan panjang menjadi grafik statistik menggunakan algoritma seperti *Naive Bayes* dan SVM. Al-Smadi et al. [20] lebih menekankan pada analisis berbasis aspek, menggunakan SVM dengan *kernel* linear untuk meningkatkan akurasi klasifikasi ulasan hotel. Nasr et al. [21] mengimplementasikan analisis sentimen pada ulasan hotel menggunakan GraphLab, memanfaatkan berbagai algoritma seperti *boosted trees, logistic regression*, dan SVM. Meskipun efektif, performa sistem menurun pada dataset besar.

Jin et al. [22] mengembangkan kerangka kerja untuk mengidentifikasi aspek fitur produk dari ulasan *online*, membantu desainer produk memahami kebutuhan pelanggan dengan lebih baik melalui teknik penambangan data dan pembelajaran mesin. Fang et al. [23] memanfaatkan metode klasifikasi seperti *Naive Bayes*, SVM, dan *random forest* untuk kategorisasi polaritas sentimen, mencapai *F1 score* tinggi pada tingkat kalimat dan ulasan. Haque et al. [24] menambahkan dimensi *active learning* dalam pelabelan dataset ulasan Amazon, mempercepat proses pembelajaran mesin dan memberikan akurasi tinggi dalam prediksi sentimen. Akhirnya, Lei et al. [25] memadukan analisis sentimen sosial dengan sistem rekomendasi, mengukur sentimen pengguna, pengaruh interpersonal, dan reputasi item berdasarkan distribusi sentimen. Hasilnya menunjukkan peningkatan akurasi prediksi rating, memperkuat pentingnya integrasi analisis sentimen dalam sistem rekomendasi berbasis teks.

Berbagai pendekatan dalam analisis data berbasis teks telah menunjukkan keefektivitasannya dalam berbagai domain, seperti bisnis, sosial, dan lingkungan. Secara keseluruhan, pendekatan-pendekatan ini menegaskan pentingnya analisis berbasis teks sebagai alat utama dalam mendukung pengambilan keputusan dan juga dapat memperkuat sistem rekomendasi.

METODE

Metode penelitian yang akan dilakukan pada penelitian ini akan melalui beberapa tahap, yaitu pengumpulan data, *preprocessing* data, pemodelan prediksi menggunakan algoritma *XGBoost Regressor*, *Random Forest Regressor, Support Vector Regressor,* KNN *Regressor*, dan terakhir adalah evaluasi, seperti pada Gambar 1.



Gambar 1. Alur Tahap Penelitian

Pengumpulan Data

Data diperoleh dari laman website *Google Scholar*. *Google Scholar* adalah mesin pencarian dari *Google* yang dapat membantu para pengguna untuk mencari sumber-sumber akademik dan literatur ilmiah. Pada laman *website* tersebut, terdapat sebuah *search bar* yang dapat digunakan untuk mencari sumber-sumber yang dibutuhkan. Untuk penelitian ini, kata kunci yang dipilih adalah kata kunci yang memiliki keterkaitan dengan topik *academic integrity* dengan rincian sebagai berikut:

1. *Academic Integrity*
2. *Contract Cheating K-12*
3. *Plagiarism Academic Integrity*
4. *Contract Cheating Higher Education*
5. *Exam Cheating Higher Education*
6. *Collusion Higher Education*
7. *Contract Cheating*
8. *Exam Cheating K-12*
9. *Research Fraud Higher Education*
10. *Research Fraud Academic Integrity*
11. *Exam Cheating Higher Education*

Setelah kata kunci dimasukkan, dilakukan proses *scraping* untuk mengambil judul, abstrak, tahun publikasi dan *link* publikasi. *Data scraping* adalah sebuah teknik yang dilakukan oleh program komputer untuk mengumpulkan informasi dari data yang dapat dibaca manusia dari sebuah *platform* [5]. Proses *scraping* dilakukan menggunakan sebuah *tool* bernama *Octoparse* untuk mengotomasi proses pengambilan informasi. Data yang dihasilkan dari proses *scraping* ini sebanyak 7121 data dengan 4 kolom seperti pada Tabel 1:

Tabel 1. Deskripsi Kolom

Nama Kolom	Deskripsi
Judul publikasi	Judul dari publikasi yang berhubungan dengan topik *academic integrity*
Abstrak publikasi	Abstrak atau ringkasan dari publikasi, yang memberikan gambaran umum tentang studi, metode, dan temuan
Tahun publikasi	Tahun terbitnya publikasi
Link publikasi	Tautan URL untuk mengakses publikasi

Preprocessing Data

Setelah data berhasil dikumpulkan, tahap selanjutnya adalah melakukan *preprocessing* data. Proses ini bertujuan untuk mengubah data teks pada kolom abstrak menjadi representasi

numerik agar dapat digunakan sebagai input oleh model pembelajaran mesin. Langkah-langkah *preprocessing* yang dilakukan adalah sebagai berikut:

1. Menambahkan kolom-kolom baru: Kolom baru yang akan dibuat adalah kolom *Query* dan *Rating* (1-5) seperti bisa dilihat pada Gambar 2. Kolom *Query* digunakan sebagai keterangan kata kunci terhadap suatu publikasi. Kata kunci hanya dimasukan pada baris awal saja.

Query	Judul publikasi	Abstrak publikasi	Tahun publikasi	Link publi	Rating (1-5)
Contract Cheating K-12	Ethics, edtech, an	... -tech with the rise	2022	https:/	4
	Academic integrity	... to academic integr	2019	https:/	4
	Interinstitutional p	... We recognize that	2019	https:/	3
	Contract cheating	... This study offers a	2022	https:/	5
	Formalising Preser	... When these K-12 s	2023	https:/	4
	Unique data sets a	... and prove contract	2022	https:/	2
	Bad apples or bad	... predictors of orgar	2020	https:/	4
	Contract Cheating	... Much of what we l	2021	https:/	4
	Multilingual essay	... of suppliers and ex	2019	https:/	3
	Administrative cor	... funds to provide e	2021	https:/	4
	Plagiarism as a soc	... A social contract re	2021	https:/	5
	AI proctoring: Aca	... These ethical conc	2022	https:/	3
	A comparison of fa	... In contract cheatir	2022	https:/	2

Gambar 2. Contoh Kolom Query

Kolom *Rating* (1-5) adalah kolom untuk menunjukkan *rating* keterkaitan abstrak terhadap kata kunci yang dipilih dengan skala 1-5, dengan penjelasan sebagai berikut:

a) 1: Sangat rendah - Abstrak memiliki sedikit atau tidak ada keterkaitan dengan kata kunci; kontennya kurang relevan.
b) 2: Rendah - Abstrak hanya mengandung sedikit elemen yang terkait dengan kata kunci, tetapi tidak sepenuhnya membahas topik.
c) 3: Cukup - Abstrak mencakup topik secara umum dan memiliki keterkaitan yang cukup dengan kata kunci.
d) 4: Tinggi - Abstrak memiliki keterkaitan yang kuat dengan kata kunci dan mencakup topik dengan baik.
e) 5: Sangat tinggi - Abstrak sangat relevan dengan kata kunci dan sepenuhnya membahas topik dengan mendalam.

2. Menghapus *missing values*. *Missing value* merupakan data yang hilang atau tidak tersedia dalam dataset. *Missing value* perlu ditangani dengan tepat agar analisis yang dilakukan tidak bias dan tetap akurat. Pada studi ini, data yang memiliki nilai kosong akan dihapus untuk menghindari dampak negatif dari keberadaan *missing value*.
3. Dalam penelitian ini, penanganan data duplikasi dilakukan dengan cara menyimpan hanya data yang pertama kali muncul dari setiap entri yang terdeteksi sebagai duplikasi. Proses ini diawali dengan mengidentifikasi duplikasi pada kolom 'Abstrak publikasi' untuk memastikan data yang dianggap duplikat memang memiliki kalimat yang sama. Setelah duplikasi teridentifikasi, entri pertama dari setiap data yang memiliki duplikasi disimpan, sementara entri-entri lainnya dihapus. Setelah langkah ini selesai, dataset diperiksa kembali untuk memastikan bahwa tidak ada lagi duplikasi yang tersisa. Pendekatan ini diterapkan untuk menjaga integritas dataset dan memastikan bahwa data yang digunakan dalam analisis tidak mengandung informasi redundan yang dapat memengaruhi hasil prediksi.

4. Memastikan *rating* hanya dalam rentang 1-5. Hal ini perlu dilakukan agar tidak tercipta *outlier* atau nilai yang tidak wajar dalam hasil analisis dikarenakan terdapat nilai yang berada di luar rentang yang ditetapkan.
5. *Stop-word Removal*: Menghapus kata-kata umum pada token yang tidak memberikan informasi penting dalam analisis teks, seperti "dan", "atau", "untuk", dan sebagainya [6].
6. *Tokenization*: Mengubah teks abstrak menjadi token-token yang lebih kecil, yaitu kata-kata individu yang akan dianalisis lebih lanjut [26].
7. *Lemmatization*: Proses untuk mengubah token menjadi bentuk kata dasar, misalnya "mencari" menjadi "cari" [27].
8. TF-IDF *(Term Frequency-Inverse Document Frequency)*: Menghitung bobot untuk setiap kata dalam setiap dokumen. TF-IDF mengukur seberapa penting sebuah kata dalam dokumen berdasarkan frekuensi kemunculannya dalam dokumen dan *corpus* secara keseluruhan [6]. Penggunaan TF-IDF diharapkan dapat memberikan pemahaman yang lebih baik terhadap relevansi setiap kata dalam abstrak terhadap topik yang dicari.

   Rumus TF-IDF dapat dilihat dalam persamaan (3.1).

$$w_{t,a} = tf_{t,a} \times idf_t \quad (3.1)$$

   Di mana $tf_{t,a}$ adalah frekuensi *term* dari *term* t dalam abstrak a, idf_t sama dengan $\log\left(\frac{A}{df_t}\right)$, di mana A adalah jumlah total abstrak yang dikumpulkan dan df_t adalah frekuensi abstrak di mana *term t* muncul dalam abstrak yang dikumpulkan. Jika *term t* sering muncul dalam abstrak a, maka nilai $tf_{t,a}$ akan meningkat; df yang rendah menghasilkan nilai idf_t yang tinggi. Bobot TF-IDF yang relatif tinggi menunjukkan bahwa term t adalah term penting dalam abstrak a [6]. TF-IDF diterapkan untuk mengubah teks ini menjadi fitur numerik yang merepresentasikan relevansi setiap *term* dalam abstrak terhadap keseluruhan dataset. Transformasi ini memungkinkan model regresi linear untuk mengkuantifikasi hubungan antara isi abstrak dan *rating* yang diberikan.

   Dengan menekankan *term* unik dalam setiap abstrak, TF-IDF memastikan bahwa model berfokus pada *term* kunci yang relevan dengan *academic integrity* daripada kata-kata yang umum atau bersifat generik.
9. Menangani distribusi rating yang tidak seimbang

   Dataset awal menunjukkan distribusi *rating* yang tidak seimbang, di mana *rating* 4 dan 3 memiliki jumlah data jauh lebih banyak dibandingkan *rating* 1, 2, dan 5. Distribusi awal data diperlihatkan pada Tabel 2.

Tabel 2. Distribusi Awal Nilai *Rating*

Rating (1-5)	Jumlah
1	886
2	876
3	1630
4	1861
5	666

Untuk menangani ketidakseimbangan ini, diterapkan metode *Synthetic Minority Oversampling Technique* (SMOTE). SMOTE bekerja dengan menghasilkan sampel sintetis untuk kelas minoritas dengan melakukan interpolasi antara sampel minoritas yang ada.

Setelah SMOTE diterapkan, distribusi data menjadi seimbang dengan jumlah sampel yang sama untuk setiap kelas, yaitu 1861 buah data.

Proses ini memastikan bahwa setiap kelas memiliki representasi yang cukup dalam model, sehingga model dapat belajar tanpa ada kecenderungan terhadap suatu nilai tertentu.

10. *Dimensionality Reduction*

Dataset hasil TF-IDF memiliki dimensi yang sangat tinggi, dengan 13,462 fitur dan 5,917 baris seperti yang ditampilkan pada Gambar 3.

	0000	00000	000001	00001	0001	0005	001	0014	0015	0018	...	актуальні	питання	in	it	タイトル	謝赫	financial	find	first	five
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
5912	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
5913	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
5914	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
5915	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
5916	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

5917 rows × 13462 columns

Gambar 3. Hasil Dimensi TF-IDF

Hal ini menimbulkan tantangan seperti meningkatnya kebutuhan komputasi secara signifikan, yang dikenal sebagai "curse of dimensionality" [28], mengurangi kemampuan generalisasi karena model dapat mempelajari noise pada data, serta membutuhkan waktu lebih lama dan memori lebih besar untuk memproses dataset besar dengan dimensi tinggi.

Untuk mengatasi permasalahan tersebut, digunakan teknik Truncated Singular Value Decomposition (SVD) untuk mereduksi dimensi data. Berdasarkan analisis terhadap varians kumulatif yang dijelaskan oleh berbagai jumlah komponen:

1. 1000 komponen menjelaskan 61,33% dari varians.
2. 1500 komponen menjelaskan 72,46% dari varians.
3. 2000 komponen menjelaskan 80,34% dari varians.
4. 3000 komponen menjelaskan 90,54% dari varians.

Dengan mempertimbangkan keseimbangan antara efisiensi komputasi dan retensi informasi, 1500 komponen dipilih untuk reduksi dimensi karena mampu menangkap lebih dari 70% varian sekaligus mempertahankan kelayakan komputasi.

Pemodelan Prediksi

Setelah melalui proses reduksi dimensi TF-IDF, hasil tersebut akan digunakan menjadi basis untuk pemodelan menggunakan empat model machine learning untuk melakukan prediksi rating berdasarkan fitur teks abastrak. Berikut adalah penjelasan lebih rinci mengenai konfigurasi dan karakteristik masing-masing model:

1. *XGBoost Regressor*

XGBoost adalah algoritma berbasis pohon keputusan yang banyak digunakan oleh *data scientist* untuk menghadapi berbagai tantangan pada *machine learning* [29]. Konfigurasi model dalam penelitian ini adalah sebagai berikut:

a. Jumlah Estimator: Model menggunakan 500 pohon keputusan (*estimators*), yang menentukan jumlah iterasi *boosting*.

b. *Learning Rate*: Nilai *learning rate* sebesar 0.05 digunakan untuk mengontrol besarnya pembaruan dalam setiap iterasi. *Learning rate* yang lebih kecil memastikan pembelajaran lebih stabil dan mengurangi risiko *overfitting*.
c. Kedalaman Maksimum Pohon *(max_depth):* Kedalaman pohon maksimum diatur menjadi 10, sehingga memungkinkan model menangkap hubungan yang lebih kompleks dalam data.
d. Optimisasi Tujuan: Model dioptimalkan untuk meminimalkan *squared error*, yang cocok untuk tugas regresi karena penalti besar diberikan pada prediksi dengan kesalahan besar.

2. *Random Forest Regressor*

Random Forest adalah algoritma *machine learning ensemble* yang dilakukan dengan cara membangun beberapa pohon keputusan yang independen dan tidak berkorelasi selama pelatihan, kemudian menggabungkan outputnya untuk menghasilkan prediksi yang lebih kuat dan akurat [30]. Konfigurasi model yang dilakukan pada penelitian ini adalah sebagai berikut:

a. Jumlah Estimator: Sebanyak 500 pohon keputusan digunakan untuk menghasilkan prediksi melalui metode rata-rata.
b. Kedalaman Maksimum Pohon *(max_depth)*: Kedalaman maksimum pohon diatur menjadi 20 untuk menghindari *overfitting* sekaligus memastikan pohon menangkap cukup banyak informasi dari data.
c. Komputasi Paralel: Penggunaan parameter *n_jobs*=-1 memungkinkan pemrosesan paralel, yang memanfaatkan semua inti CPU untuk mempercepat pelatihan dan prediksi.
d. *Random Forest* bekerja dengan baik untuk data berdimensi tinggi dan dataset besar, meskipun kinerjanya sedikit menurun pada data hasil reduksi dimensi seperti dalam penelitian ini.

3. *Support Vector Regressor (SVR)*

SVR adalah metode regresi yang memanfaatkan margin kesalahan epsilon untuk menghasilkan prediksi yang akurat tanpa terlalu sensitif terhadap *outlier*. Pengaturan model meliputi:

1. Kernel RBF: Kernel *Radial Basis Function* (RBF) digunakan untuk menangkap hubungan non-linear dalam data, sehingga menjadikannya sangat efektif pada dataset dengan pola yang kompleks.
2. *Hyperparameter* C: Nilai C diatur menjadi 10, yang mengontrol tingkat regularisasi. Nilai C yang lebih besar mengurangi *margin* kesalahan tetapi meningkatkan risiko *overfitting*.
3. Epsilon: Nilai epsilon sebesar 0.1 digunakan untuk menentukan *margin* toleransi di mana prediksi tidak dihitung sebagai kesalahan. Hal ini memberikan keseimbangan antara generalisasi dan akurasi.

4. *K-Nearest Neighbors Regressor (KNN Regressor)*

KNN adalah model berbasis *instance* yang memprediksi nilai target berdasarkan rata-rata nilai dari tetangga terdekat dalam ruang fitur. Konfigurasi model meliputi:

1. Jumlah Tetangga (*Neighbors*): Sebanyak 5 tetangga digunakan untuk menentukan prediksi. Pemilihan jumlah tetangga memengaruhi keseimbangan antara bias dan variansi.
2. Bobot Jarak: Pemberian bobot berdasarkan jarak diterapkan, sehingga tetangga yang lebih dekat memiliki pengaruh lebih besar terhadap prediksi dibandingkan tetangga yang lebih jauh.
3. KNN mudah dipahami dan diterapkan, tetapi sensitif terhadap metrik jarak dan parameter jumlah tetangga, terutama pada data berdimensi tinggi.

Evaluasi Model

Setelah pelatihan selesai, model-model algoritma tersebut akan dievaluasi dengan menggunakan metrik sebagai berikut:

1. *Mean Absolute Error (MAE)*

   Mengukur rata-rata kesalahan absolut antara nilai prediksi dan nilai aktual [31]. Rumus MAE dapat dilihat dalam persamaan (3.2).

$$MAE = \frac{1}{n}\sum_{k=1}^{n}\left|Y_k - \hat{Y}_k\right| \quad (3.2)$$

2. *Mean Squared Error (MSE)*

   Mean Squared Error adalah adalah sebuah metode pengukuran dengan cara menghitung rata-rata kuadrat perbedaan nilai prediksi dan nilai sebenarnya. Nilai MSE yang rendah mengindikasikan bahwa nilai prediksi mendekati nilai yang sebenarnya [32]. MSE banyak digunakan pada model-model regresi [33]. Rumus MSE dapat dilihat dalam persamaan (3.3) [32].

$$MSE = \frac{1}{n}\sum_{k=1}^{n}(Y_k - \hat{Y}_k)^2 \quad (3.3)$$

   Keterangan:

   Y_k : nilai *rating* sebenarnya
   $\hat{Y}_k$: *rating* prediksi
   n : jumlah prediksi

3. *Coefficient of Determination (R^2)*

   R^2 mengukur seberapa besar variasi (atau perubahan) pada variabel dependen (yang ingin diprediksi) dapat dijelaskan oleh variabel independen (faktor-faktor yang digunakan untuk membuat prediksi). Rumus R^2 dapat dilihat pada persamaan (3.4).

$$R^2 = 1 - \frac{\sum_{i=1}^{m}(X_i - Y_i)^2}{\sum_{i=1}^{m}\left(\hat{Y}_i - Y_i\right)^2} \quad (3.4)$$

   Nilai R^2 berada pada rentang (0, 1), jika R^2 bernilai 1, maka *independent variable* memiliki keterkaitan terhadap variabel dependen, bila 0 maka tidak memiliki keterkaitan [34].

4. *Accuracy*

   Selain menggunakan MAE, MSE, dan R^2 sebagai metrik evaluasi, penelitian ini juga mengevaluasi kinerja model dalam hal akurasi pengelompokkan relevansi abstrak. Dalam hal ini, *rating* yang diprediksi oleh model dikategorikan ke dalam tiga kelompok relevansi sebagai berikut:

   1. Relevan: *Rating* 4 atau 5.
   2. Netral: *Rating* 3.
   3. Tidak Relevan: *Rating* 1 atau 2.

Proses penghitungan akurasi dilakukan dengan langkah-langkah berikut:

1. Prediksi *rating* yang dihasilkan oleh model dibulatkan ke nilai *integer* terdekat (1-5) dan dikelompokkan ke dalam tiga kategori di atas.
2. Rating aktual juga dikelompokkan ke dalam kategori relevansi yang sama.
3. Akurasi dihitung dengan membandingkan kesesuaian antara kelompok prediksi dengan kelompok aktual menggunakan persamaan (3.5).

$$Akurasi = \frac{Jumlah\ Prediksi\ yang\ Sesuai}{Total\ Jumlah\ Data\ Uji} \quad (3.5)$$

Hasil akurasi ini memberikan gambaran tambahan mengenai kemampuan model untuk tidak hanya memprediksi nilai rating, tetapi juga mengelompokkan relevansi abstrak secara kualitatif. Evaluasi ini penting untuk memahami sejauh mana model dapat mendukung pengambilan keputusan berbasis relevansi abstrak pada literatur ilmiah.

HASIL DAN PEMBAHASAN

Kinerja dari empat model *machine learning* yang dievaluasi dalam penelitian ini dirangkum pada Tabel 3. Secara umum, *Support Vector Regressor* (SVR) merupakan model terbaik jika dibandingkan dengan model-model lainnya. Model tersebut dapat mencari dan mengidentifikasi pola yang cukup kompleks dalam data [20]. Model ini dapat menghasilkan akurasi tinggi dengan derajat *error* yang rendah. SVR mencatat nilai *Mean Absolute Error* (MAE) terendah sebesar 0.4980, *Mean Squared Error* (MSE) terendah sebesar 0.5237, dan R^2 *Score* tertinggi sebesar 0.7321, serta akurasi pengelompokan relevansi tertinggi sebesar 72.35%. Hal ini menunjukkan bahwa SVR tidak hanya mampu menghasilkan prediksi yang paling akurat secara numerik, tetapi juga memiliki kemampuan terbaik dalam mengelompokkan abstrak berdasarkan tingkat relevansinya.

Model KNN berada di posisi kedua dalam hal akurasi dengan nilai 63.15%, meskipun nilai MAE, MSE, dan R^2-nya tidak lebih baik dibandingkan *XGBoost*, yaitu MAE 0.6704, MSE 0.8246 dan R^2 0.5782. Hal ini menunjukkan bahwa KNN lebih unggul dalam pengelompokan berbasis jarak untuk relevansi abstrak, terutama pada dataset dengan pola lokal yang lebih sederhana.

Tabel 3. Hasil Evaluasi Metrik

Model	*Mean Absolute Error* (MAE)	*Mean Squared Error* (MSE)	R^2 *Score*	*Accuracy*
XGBoost Regressor	0.6650	0.7524	0.6151	62.29%
Random Forest Regressor	0.7435	0.8370	0.5718	57.8%
Support Vector Regressor	0.4980	0.5237	0.7321	72.35%
K-Nearest Neighbors Regressor (KNN)	0.6704	0.8246	0.5782	63.15%

Model *XGBoost Regressor* menunjukkan kinerja yang kompetitif dengan akurasi sebesar 62.29%, nilai MAE sebesar 0.6650, MSE sebesar 0.7524, dan R^2 sebesar 0.6151. Meskipun lebih unggul dalam regresi numerik dibandingkan KNN, *XGBoost* kurang optimal dalam pengelompokan relevansi abstrak.

Random Forest Regressor mencatat kinerja terendah dengan nilai akurasi sebesar 57.88%, serta nilai MAE tertinggi sebesar 0.7435, MSE sebesar 0.8370, dan R^2 terendah sebesar 0.5718. Model ini mengalami kesulitan dalam menangkap pola yang lebih kompleks dalam dataset ini.

Proses *preprocessing*, yang mencakup penghapusan duplikasi data, penerapan TF-IDF untuk representasi teks, penanganan ketidakseimbangan data menggunakan SMOTE, dan reduksi dimensi melalui *Truncated* SVD, terbukti efektif dalam meningkatkan efisiensi dan akurasi model. *Truncated* SVD berhasil mengurangi dimensi data secara signifikan dengan tetap mempertahankan lebih dari 70% variansi awal, yang memungkinkan model untuk bekerja dengan fitur yang lebih ringkas tanpa kehilangan informasi penting.

Berikut penjelasan hasil berdasarkan masing-masing metrik:

1. *Mean Absolute Error (MAE):*
   a. SVR memiliki nilai MAE terendah sebesar 0.4980, menunjukkan bahwa model ini menghasilkan prediksi yang paling dekat dengan nilai aktual secara rata-rata.
   b. *XGBoost* berada di urutan kedua dengan MAE sebesar 0.6650, menghasilkan prediksi yang cukup akurat, meskipun kurang baik dibandingkan SVR.
   c. KNN memiliki MAE sebesar 0.6704, sedikit lebih tinggi dari *XGBoost*, menunjukkan kesalahan rata-rata yang lebih besar.
   d. *Random Forest* memiliki MAE tertinggi sebesar 0.7435, menunjukkan kesalahan rata-rata yang paling besar dibandingkan model lainnya.
2. *Mean Squared Error (MSE):*
   a. SVR memiliki nilai MSE terendah sebesar 0.5237, menunjukkan bahwa model ini menghasilkan prediksi yang konsisten dengan sedikit penyimpangan besar.
   b. *XGBoost* berada di posisi kedua dengan MSE sebesar 0.7524, yang masih menunjukkan konsistensi prediksi yang baik, tetapi tidak sebaik SVR.
   c. KNN memiliki nilai MSE sebesar 0.8246, menunjukkan bahwa model ini menghasilkan lebih banyak kesalahan besar dibandingkan dua model sebelumnya.
   d. *Random Forest* memiliki MSE tertinggi sebesar 0.8370, menunjukkan prediksi yang paling tidak konsisten di antara keempat model.
3. *R^2 Score:*
   a. SVR memiliki nilai R^2 tertinggi sebesar 0.7321, menunjukkan bahwa model ini mampu menjelaskan 73.21% variabilitas dalam data, menjadi model paling andal untuk tugas ini.
   b. *XGBoost* memiliki R^2 sebesar 0.6151, yang menunjukkan kemampuan penjelasan variabilitas data yang cukup baik, meskipun tidak sebaik SVR.
   c. KNN dengan nilai R^2 sebesar 0.5782 menunjukkan bahwa model ini masih dapat menjelaskan lebih dari setengah variabilitas data, tetapi kurang akurat dibandingkan dua model sebelumnya.
   d. *Random Forest* memiliki nilai R^2 terendah sebesar 0.5718, menunjukkan bahwa model ini memiliki kemampuan paling rendah dalam menjelaskan variabilitas data.
4. *Accuracy*:
   a. SVR memiliki akurasi tertinggi sebesar 72.35%, yang menunjukkan bahwa model ini paling efektif dalam memprediksi relevansi abstrak secara konsisten. Keunggulan ini sesuai dengan nilai MAE, MSE, dan R^2 yang juga terbaik di antara semua model, mengindikasikan bahwa SVR tidak hanya menghasilkan prediksi yang akurat secara numerik, tetapi juga mampu mengelompokkan prediksi tersebut dengan baik.
   b. KNN memiliki akurasi sebesar 63.15%, berada di posisi kedua setelah SVR. Hasil ini menunjukkan bahwa meskipun nilai MAE, MSE, dan R^2 dari KNN tidak sebaik XGBoost, model ini lebih unggul dalam mengelompokkan relevansi abstrak. Kemampuan ini kemungkinan dipengaruhi oleh pendekatan berbasis jarak yang digunakan oleh KNN, yang memungkinkan model ini menangkap pola lokal dalam data.

c. *XGBoost Regressor* memiliki akurasi sebesar 62.29%, sedikit lebih rendah dari KNN. Meskipun *XGBoost* memiliki nilai R^2 yang lebih tinggi dibandingkan KNN, akurasi pengelompokannya lebih rendah, yang mengindikasikan bahwa model ini cenderung menghasilkan prediksi numerik yang lebih baik tetapi kurang optimal dalam mengelompokkan relevansi abstrak.
d. *Random Forest* memiliki akurasi terendah sebesar 57.88%, mengonfirmasi bahwa model ini mengalami kesulitan dalam mengelompokkan relevansi abstrak dengan baik. Hal ini juga konsisten dengan nilai MAE, MSE, dan R^2 model ini yang menunjukkan kinerja paling rendah dibandingkan model lainnya.

KESIMPULAN DAN SARAN

Penelitian ini mengevaluasi empat algoritma *machine learning*, yaitu *XGBoost Regressor, Random Forest Regressor, Support Vector Regressor (SVR), dan K-Nearest Neighbors Regressor (KNN)* untuk memodelkan prediksi *rating* abstrak publikasi ilmiah. Hasil evaluasi menunjukkan bahwa SVR memiliki performa terbaik, disusul oleh KNN, *XGBoost,* dan *Random Forest*. *Preprocessing* dapat meningkatkan kinerja dari model.

Dari hasil analisis penelitian, kesimpulan yang dapat diambil adalah SVR merupakan model terbaik untuk tugas prediksi *rating* abstrak berdasarkan relevansi, sementara model lainnya memberikan kinerja yang kompetitif tergantung pada keunggulan masing-masing. Untuk penelitian selanjutnya, disarankan untuk mengeksplorasi kombinasi model *ensemble* atau pendekatan *hybrid* guna meningkatkan akurasi prediksi, serta menerapkan metode ini pada dataset yang lebih besar dan beragam untuk menguji generalisasi model secara lebih luas. Kami juga tertarik melakukan *topic modeling* [35] dan *clustering* [36] guna mendapatkan pola data yang lebih komprehensif.

DAFTAR PUSTAKA

[1] S. H. S. University,"Google Scholar Tutorial," [Online]. Available: https://library.shsu.edu/research/guides/tutorials/googlescholar/index.html. [Accessed 19 December 2024].

[2] R. Vine, "Google Scholar," *J Med Libr Assoc,* vol. 94, no. 1, pp. 97-99, 2006.

[3] M. S. Tullu, "Writing the title and abstract for a research paper Being concise, precise, and meticulous is the key," *Saudi Journal of Anaesthesia,* vol. 13, no. Suppl 1, pp. S12 - S17, 2019.

[4] B. Butzer, "Bias in the evaluation of psychology studies: A comparison of parapsychology versus neuroscience," *EXPLORE,* vol. 16, no. 6, pp. 382-391, 2020.

[5] P. Yang, X. Sun, W. Li and S. Ma, "Automatic academic paper rating based on modularized hierarchical convolutional neural network," *arXiv preprint arXiv:1805.03977,* 2018.

[6] L. T. K. Nguyen, H. H. Chung, K. V. Tuliao and T. M. Lin, "Using XGBoost and skip-gram model to predict online review popularity.," *SAGE Open,* vol. 10, no. 4, 2020.

[7] N. Ashgar, "Yelp Dataset Challenge: Review Rating Prediction," *arXiv preprint arXiv,* 2016.

[8] V. S. Anoop, T. K. A. Krishnan, A. Daud, A. Banjar and A. Bukhari, "Climate Change Sentiment Analysis Using Domain Specific Bidirectional Encoder Representations From Transformers," *IEEE Access,* vol. 12, pp. 114912-114922, 2024.

[9] S. EYVAZI-ABDOLJABBAR, S. KIM, M.-R. FEIZI-DERAKHSHI, Z. FARHADI and D. A. MOHAMMED, "An Ensemble-Based Model for Sentiment Analysis of Persian Comments on Instagram," *IEEE Access 12,* pp. 151223-151235, 2024.

[10] P. Ghosh, O. Samanta, T. Goto and S. Sen, "Sales Forecasting of Overrated Products: Fine Tuning of Customer's Rating by Integrating Sentiment Analysis," *IEEE Access,* vol. 12, pp. 69578-69592, 2024.

[11] D. Amangeldi, A. Usmanova and P. Shamoi, "Understanding Environmental Posts: Sentiment and Emotion Analysis of Social Media Data," *IEEE Access,* vol. 99, pp. 1-1, 2024.

[12] N. R. Ram, S. Gautum, A. Jadeja and H. Joisar, "Social Media Sentiment Analysis Using Twitter Dataset," in *2024 1st International Conference on Cognitive, Green and Ubiquitous Computing (IC-CGU)*, 2024.

[13] S. Raheja and A. Asthana, "Sentiment Analysis of Tweets During the COVID-19 Pandemic Using Multinomial Logistic Regression," *International Journal of Software Innovation,* vol. 11, no. 1, 2023.

[14] Y. Xu and N. F. Ibrahim, "Cross-Domain Aspect-Based Sentiment Analysis for Enhancing Customer Experience in Electronic Commerce," *Adv. Artif. Intell. Mach. Learn.,* vol. 4, no. 3, pp. 2593-2613, 2024.

[15] R. Herrero-Álvarez, E. Callejas-Castro, G. Miranda and C. León, "Analysis of Sentiment Toward Computer Science in Pre-University Education," *IEEE Access,* vol. 12, pp. 71205-71218, 2024.

[16] W.-K. Chen, C. Lin and Y.-S. Tai, "Text-Based Rating Predictions on Amazon Health & Personal Care Product Review," in *28th Modern Artificial Intelligence and Cognitive Science Conference (MAICS 2017)*, 2015.

[17] H. Isah, P. Trundle and D. Neagu, "Social media analysis for product safety using text mining and sentiment analysis," *2014 14th UK Workshop on Computational Intelligence (UKCI),* pp. 1-7, 2014.

[18] S. Mukherjee and P. Bhattacharyya, "Feature Specific Sentiment Analysis for Product Reviews," in *Gelbukh, A. (eds) Computational Linguistics and Intelligent Text Processing*, 2012.

[19] A. Bhatt, A. Patel, H. Chheda and K. Gawande, "Amazon Review Classification and Sentiment," *(IJCSIT) International Journal of Computer Science and Information Technologies,,* vol. 6, no. 6, pp. 5107-5110, 2015.

[20] M. AL-Smadi, O. Qwasmeh, B. Talafha, M. Al-Ayyoub, Y. Jararweh and E. Benkhelifa, "An enhanced framework for aspect-based sentiment analysis of Hotels' reviews: Arabic reviews case study," in *2016 11th International Conference for Internet Technology and Secured Transactions (ICITST)*, Barcelona, 2016.

[21] M. M. Nasr and E. Shaaban, "Building Sentiment analysis Model using Graphlab," *International Journal of Scientific and Engineering Research ,* vol. 8, no. 6, pp. 1155-1160, 2017.

[22] J. Jin and P. Ji, "Mining online product reviews to identify consumers' fine-grained concerns," in *12th International Symposium on Operations Research and its Applications in Engineering, Technology and Management (ISORA 2015)*, Luoyang, 2015.

[23] X. Fang and J. Zhan, "Sentiment analysis using product review data," *Journal of Big Data,* vol. 2, no. 5, 2015.

[24] T. U. Haque, N. N. Saber and F. M. Shah, "Sentiment analysis on large scale Amazon product reviews," in *2018 IEEE International Conference on Innovative Research and Development (ICIRD)*, Bangkok, 2018.

[25] X. Lei, X. Qian and G. Zhao, "Rating Prediction Based on Social Sentiment From Textual Reviews," *IEEE Transactions on Multimedia,* vol. 18, no. 9, pp. 1910-1921, 2016.

[26] G. Grefenstette, "Tokenization," in *van Halteren, H. (eds) Syntactic Wordclass Tagging. Text, Speech and Language Technology*, Springer, 1999.

[27] J. Plisson, N. Lavrac and D. Mladenic, "A rule based approach to word lemmatization," in *Proceedings of IS*, 2004.

[28] R. Nisbet, G. Miner and K. Yale, "Chapter 5 - Feature Selection,," in *Handbook of Statistical Analysis and Data Mining Applications (Second Edition)*, Academic Press, 2018, pp. 89-97.

[29] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system.," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016.

[30] Z. E. Mrabet, N. Sugunaraj, P. Ranganathan and S. Abhyankar, "Random Forest Regressor-Based Approach for Detecting Fault Location and Duration in Power Systems," *Sensors,* vol. 22, no. 2, p. 458, 2022.

[31] A. S. Rajawat, O. Mohammed, R. N. Shaw and A. Ghosh, "Chapter six - Renewable energy system for industrial internet of things model using fusion-AI," *Applications of AI and IOT in Renewable Energy,* pp. 107-128, 2022.

[32] Z. Zhao, L. Alzubaidi, J. Zhang, Y. Duan and Y. Gu, "A comparison review of transfer learning and self-supervised learning: Definitions, applications, advantages and limitations," *Expert Systems with Applications,* vol. 242, no. 122807, 2024.

[33] K. Tyagi, C. Rane, Harshvardhan and M. Manry, "Chapter 4 - Regression analysis," in *Artificial Intelligence and Machine Learning for EDGE Computing*, Academic Press, 2022, pp. 53-63.

[34] D. Chicco, M. J. Warrens and G. Jurman, "The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation," *PeerJ Comput Sci,* 2021.

[35] C. B. Pavithra, and J. Savitha. Topic Modeling for Evolving Textual Data Using LDA HDP NMF BERTOPIC and DTM With a Focus on Research Papers. *Journal of Technology and Informatics (JoTI)*, vol. 5 no. 2, pp. 53-63. 2024.

[36] A. Salsabiela, A. P. Kuncoro, P. Subarkah, and P. Arsi. Rekomendasi Restock Barang di Toko Pojok UMKM Menggunakan Algoritma K-Means Clustering. *Journal of Technology and Informatics (JoTI)*, vol. 5 no. 2, pp. 87-92. 2024.