Perbandingan Peramalan Penerimaan Calon Mahasiswa Baru menggunakan Naïve Bayes dan Support Vector Machine

Parasian D.P Silitonga[1], Doni El Rezen Purba[2], Alex Rikki[3]

[1,2,3]Fakultas Ilmu Komputer, Universitas Katolik Santo Thomas, Jalan Setia Budi No.479 F, Tanjung Sari Medan, Indonesia.

ARTICLE INFORMATION

Received: September 09, 2024
Revised: September 12, 2024
Available online: Oktober 30, 2024

KEYWORDS

Support Vector Machine, Naïve Bayes, Data Mining, Peramalan, Perbandingan.

CORRESPONDENCE

Phone: +62 81327735016
E-mail: parasianirene@gmail.com

A B S T R A C T

Penelitian ini dilakukan untuk melihat perbandingan peramalan penerimaan calon mahasiswa baru menggunakan Naïve Bayes dan Support Vector Machine. Analisis menunjukkan bahwa model Naïve Bayes menghasilkan akurasi moderat sebesar 50% pada data uji dan memprediksikan jumlah pendaftar tetap berada pada kategori Tinggi dengan estimasi rata-rata sekitar 1542,5. Akurasi yang dihasilkan tergolong cukup rendah, tetapi Naïve Bayes dapat bekerja dengan baik pada prediksi berbasis kategori. Sebaliknya, model SVM yang diterapkan dalam bentuk Support Vector Regression (SVR) juga menunjukkan akurasi 50%, namun memberikan prediksi numerik yang lebih rinci, dengan estimasi jumlah pendaftar tetap sebesar 1883. SVM menunjukkan potensi yang lebih besar dalam menangani data dengan pola tren yang meningkat. Perbandingan antara kedua metode ini menunjukkan bahwa Naïve Bayes lebih cocok untuk prediksi kategori, sedangkan SVM lebih tepat untuk prediksi numerik yang lebih akurat.

PENDAHULUAN

Data mining merupakan suatu proses dalam mengumpulkan data lalu digunakan dengan cara diolah guna mengekstrak segala sesuatu informasi relevan. Akan tetapi, proses tersebut secara otomatis dapat dilakukan menggunakan perangkat lunak dengan pertolongan perhitungan statistika, matematika, ataupun teknologi Kecerdasan Buatan atau dalam bahasa Inggris disebut sebagai Artificial Intelligence/AI. Data mining sering juga disebut Knowledge Discovery in Database atau KDD [1]. Salah satu aplikasi utama dari data mining adalah penggunaannya dalam peramalan - proses memprediksi keadaan yang akan terjadi di masa depan berdasarkan data historis yang telah ada [2]. Peramalan ini sangat penting untuk mengurangi ketidakpastian dan membantu pengambil keputusan merencanakan kebijakan yang berorientasi pada masa depan.

Teknik data mining, melalui analisis data masa lalu, memungkinkan untuk menemukan pola dan kecenderungan yang berguna dalam kumpulan data besar [3]. Data mining adalah ilmu yang mempelajari cara mengumpulkan, membersihkan, mengolah, dan menganalisis data untuk memperoleh wawasan yang dapat digunakan untuk berbagai keperluan. Salah satu metode dalam data mining yang digunakan untuk memprediksi adalah klasifikasi, yaitu metode yang bertujuan untuk memprediksi kategori atau kelas berdasarkan data yang ada [4].

Ada beberapa jenis praktik data mining yang dalam kenyataannya memiliki tingkat akurasi yang tinggi dalam prediksi. Dua diantaranya penerapan metode yang sering digunakan adalah Naive Bayes dan Support Vector Machine. SVM dapat dikategorikan sebagai algoritma linier yang sering kali banyak dipakai dalam klasifikasi, regresi, estimasi kerapatan, deteksi kebaruan, dan banyak aplikasi lainnya. Pada kasus klasifikasi dua kelas, SVM berusaha mencari hyperplane yang memisahkan kedua kelas dengan margin selebar mungkin, yang berkontribusi pada akurasi generalisasi yang baik pada data yang tidak terlihat, serta mendukung pengoptimalan khusus yang memungkinkan SVM belajar dari sejumlah besar data [5].

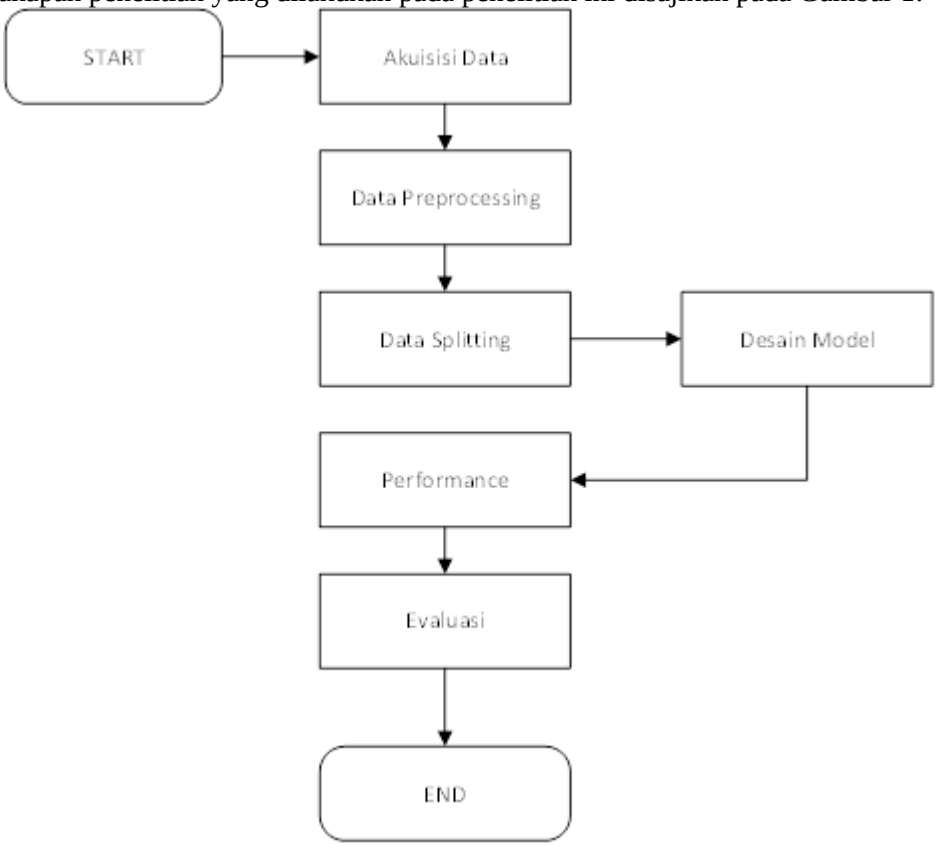
Sementara itu, Naive Bayes adalah algoritma berbasis probabilitas menggunakan Teorema Bayes, di mana fitur-fitur dalam data dianggap tidak saling bergantung. Meskipun sederhana, Naive Bayes efektif dalam mengolah data dengan cepat, bahkan jika jumlah data terbatas [6].

Mahasiswa sebagai peserta didik di perguruan tinggi memiliki peran yang sangat penting dalam mendukung pencapaian tujuan pembangunan nasional. Perguruan tinggi sebagai lembaga pendidikan memiliki tugas untuk mempersiapkan mahasiswa sesuai dengan tujuan pendidikan tinggi melalui Tridharma Perguruan Tinggi, yang meliputi pendidikan, penelitian dan pengabdian kepada masyarakat. Jumlah mahasiswa baru yang diterima pada setiap tahun ajaran merupakan salah satu faktor perencanaan dalam sistem pendidikan tinggi. Prediksi perlu dilakukan dalam jumlah mahasiswa baru tersebut untuk determinasi penggunaan fasilitas pendidikan, seperti kelas dan lain-lainnya [7].

Berdasarkan latar belakang tersebut, penelitian ini bertujuan untuk memprediksi jumlah penerimaan mahasiswa baru dengan menggunakan metode Naive Bayes dan Support Vector Machine serta membandingkan keakuratan antara kedua metode tersebut.

METODE PENELITIAN

Metode yang digunakan pada penelitian ini adalah dengan menggunakan metode Naïve Bayes dan Support Vector Machine (SVM), dimana kedua metode ini merupakan bagian dari machine learning. Sumber data yang digunakan pada penelitian ini adalah berasal dari data penerimaan calon mahasiswa baru Universitas Katolik Santo Thomas Tahun Akademik 2018/2019 sampai dengan Tahun Akademik 2023/2023. Tahapan penelitian yang dilakukan pada penelitian ini disajikan pada Gambar 1.



Gambar 1. Tahapan Penelitian

1. Pengumpulan Data

Pengumpulan data (Data mining) adalah metode yang dapat membantu para penggunanya untuk mengakses data yang besar dalam waktu yang relatif singkat. Dengan kata lain, data mining adalah alat dan aplikasi yang menggunakan analisis statistik pada data melalui suatu peoses ekstraksi atau penggalian data dan informasi yang sebelumnya tidak diketahui. Data mining adalah proses mengekstraksi data untuk menemukan informasi terbaru dengan mengacari pola atau aturan tertentu dari sejumlah data yang sangat besar, sehingga aktivitas penambangan adalah memeriksa database yang berukuran besar untuk menemukan

pola atau bentuk yang baru sehingga berguna dalam proses pengambilan keputusan [13]. Data yang diambil untuk penelitian ini berasal dari ulasan pengguna di google play store mengenai aplikasi LinkedIn. Akuisisi data merupakan tahap mengumpulkan data yang akan diolah. Data dapat diperoleh melalui berbagai metode seperti wawancara langsung, kuisioner, observasi atau pengambilan data dari sumber yang ada. Metode yang digunakan untuk memperoleh data harus konsisten dengan tujuan penelitian dan asumsi yang telah ditentukan sebelumnya. Data juga dapat diperoleh dengan cara mengambil data dari sumber data. Pada tahapan ini proses akuisisi data dilakukan dengan mengumpulkan data dari sistem penerimaan mahasiswa baru Universitas Katolik Santo Thomas.

2. Preprocessing

Prepocessing merupakan langkah mengubah data mentah menjadi format yang dapat dipahami adalah bagian dari teknik penambangan data. Masalah seperti data noisy, redundansi, dan kehilangan nilai dapat diatasi dengan pra-pemrosesan data [18]. Cleaning, case folding, tokenizing, stopword, stemming, dan formalisasi adalah beberapa bagian dari tahap ini menurut [12].

a. Cleaning: Proses membersihkan data dari elemen yang tidak penting, seperti kesalahan tata bahasa, konjungsi, tanda baca, dan konten asing.
b. Case folding: metode yang mengubah huruf besar pada kata atau kalimat menjadi huruf kecil.
c. Tokenizing: proses membagi atau memecah frase kata-kata yang awalnya terbentuk menjadi kata-kata individual.
d. Stopword: Seperti kata sambung dan kata ganti orang, istilah-istilah yang tidak memiliki data yang diperlukan akan dihapus selama proses ini.
e. Stemming: proses dengan setiap kata bebas dari imbuhan.

3. Term Weighting

Term weighting adalah sebuah metode pembobotan kata (term) untuk memberikan sebuah bobot atau nilai untuk kata (term) yang terkandung dalam sebuah dokumen. Bobot nilai ini menjadi ukuran besarnya jumlah dan tingkat kontribusi sebuah kata (term) untuk penentuan suatu kelas atau kategori dalam suatu dokumen. Terdapat beberapa metode pembobotan kata (term weighting) diantaranya adalah TF, TF-IDF, WIDF, dan TF-RF [6].

3.1 TF-IDF

TF (Term Frequency) adalah frekuensi dari kemunculan sebuah term dalam dokumen yang bersangkutan. IDF (Inverse Document Frequency) merupakan sebuah perhitungan dari bagaimana term didistribusikan secara luas pada koleksi dokumen yang bersangkutan. Semakin besar jumlah kemunculan suatu term (TF tinggi) dalam dokumen, semakin besar pula bobotnya atau akan memberikan nilai kesesuaian yang semakin besar [7].

4. Support Vectore Machine (SVM)

Support vector machines (SVM) merupakan algoritma linier yang dapat digunakan padaproses klasifikasi, regresi, estimasi kerapatan, deteksi kebaruan, dan aplikasi lainnya. Dalam kasus klasifikasi dua kelas yang paling sederhana, SVM memiliki hyperplane yang memisahkan dua kelas data dengan selebar margin sebisa mungkin. mengarah pada akurasi generalisasi yang baik pada data yang tidak terlihat, dan mendukung metode pengoptimalan khusus yang memungkinkan SVM untuk belajar dari sejumlah besar data [5].

Prinsip dasar SVM adalah linear classifier, dan selanjutnya dikembangkan agar dapat bekerja pada problem on-linear. dengan memasukkan konsep kernel trick pada ruang kerja berdimensi tinggi. Support Vector Machine (SVM) merupakan salah satu metode klasifikasi dengan menggunakan machine learning (supervised learning) yang memprediksi kelas berdasarkan model atau pola dari hasil proses training. Klasifikasi dilakukan dengan mencari hyperplane atau garis pembatas (decision boundary) yang memisahkan antara suatu kelas dengan kelas lain, yang dalam kasus ini garis tersebut berperan memisahkan data bersentimen positif (berlabel +) dengan data bersentimen negatif (berlabel -) [16].

5. Naïve Bayes

Naïve bayes merupakan sebuah metode klasifikasi dengan probabilitas sederhana yang mengaplikasikan Teorema Bayes dengan asumsi ketidaktergantungan (independen) yang tinggi. Penggunaan metode naive bayes pada penelitian ini didasarkan pada banyaknya dataset yang dipakai sehingga membutuhkan suatu metode yang mempunyai performansi yang cepat dalam pengklasifikasian serta keakuratan yang cukup tinggi [17].

Naive bayes adalah salah satu algoritma yang digunakan untuk klasifikasi teks serta merupakan metode machine learning yang menggunakan perhitungan probabilitas dan statistik yang dikemukakan oleh

Thomas Bayes. Algoritma tersebut digunakan untuk memprediksi probabilitas di masa depan berdasarkan pengalaman di masa lalu. Keuntungan penggunaan naive bayes adalah metode ini hanya membutuhkan jumlah data pelatihan (training data) yang kecil untuk menentukan estimasi parameter yang diperlukan dalam proses pengklasifikasian [18].

5. Evaluasi Model

Evaluasi model adalah proses menganalisis kinerja model machine learning dengan menggunakan berbagai metrik/parameter. Evaluasi model sangat penting dalam menilai efektivitas sebuah model selama fase penelitian awal dan juga berperan dalam pemantauan model kedepannya. Hasil dari algoritma pembelajaran perlu dinilai dan dianalisis dengan benar, sehingga dapat mengevaluasi performa algoritma pembelajaran yang berbeda. Performa klasifikasi diwakili oleh nilai skalar dalam metrik yang berbeda seperti accuracy, sensitivity, dan specifity [9]. Penulis menggunakan metrik evaluasi standar untuk mengukur performa model dengan menggunakan metrik True Positive, True Negative, False Positive, dan False Negative dimana Sebuah true positive adalah hasil di mana model dengan benar memprediksi kelas positif. Demikian pula, sebuah true negative adalah hasil di mana model dengan benar memprediksi kelas negatif. Sebuah false positive adalah hasil di mana model dengan salah memprediksi kelas positif. Dan sebuah false negative adalah hasil di mana model dengan salah memprediksi kelas negatif. Dengan menggunakan metrik - metrik tersebut penulis dapat mengkalkulasikan akurasi, recall, presisi, dan F1-score menggunakan persamaan 1, 2, 3, dan 4 [24].

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$Recall = \frac{TP}{TP + FN} \quad (2)$$

$$Precision = \frac{TP}{TP + FP} \quad (3)$$

$$F1 - Score = \frac{2 * TP}{2 * TP + FP + FN} \quad (4)$$

Pada persamaan (1) yaitu accuracy seberapa sering model / mesin membuat prediksi yang benar. Persamaan (2) recall menunjukan perbandingan antara True Positive (TP) dengan banyaknya data yang sebenarnya positif sedangkan precision yang ditunjukan pada persamaan (3) mengukur seberapa banyak prediksi mesin positif yang sebenarnya positif. Terakhir, F1 score yang ditunjukan pada persamaan (4) menggabungkan precision dan recall dalam satu metrik, F1 score sangat penting untuk menentukan performa model machine learning [10].

HASIL DAN PEMBAHASAN

Pada bagian ini akan membahas mengenai hasil peramalan dengan menggunakan metode Naïve Bayes dan Support Vector Machine, dalam memprediksi jumlah pendaftaran tetap untuk tahun 2025. Dataset yang digunakan dalam penelitian ini berisi informasi jumlah pendaftar awal dan jumlah pendaftar tetap selama beberapa tahun akademik yang diamna datanya diambil dari PMB dari tahun 2017 sampai 2024, serta kategori jumlah pendaftar tetap yang dikelompokkan ke dalam tiga kelas (rendah, sedang, dan tinggi). Variabel utama dalam dataset meliputi:

a. Jumlah Pendaftar Awal: Total pendaftar yang mendaftar di awal tahun akademik.
b. Jumlah Pendaftar Tetap: Total pendaftar yang diterima atau tetap setiap tahun.
c. Tahun Ajaran: Tahun akademik yang menjadi dasar data.
d. Kategori Pendaftar Tetap: Kategori berdasarkan jumlah pendaftar tetap, yaitu:
    1. Rendah: < 800
    2. Sedang: 800–1200
    3. Tinggi: > 1200

    Dengan data sebagai berikut:

Tabel 1. Data mahasiswa tahun 2017 sampai 2024

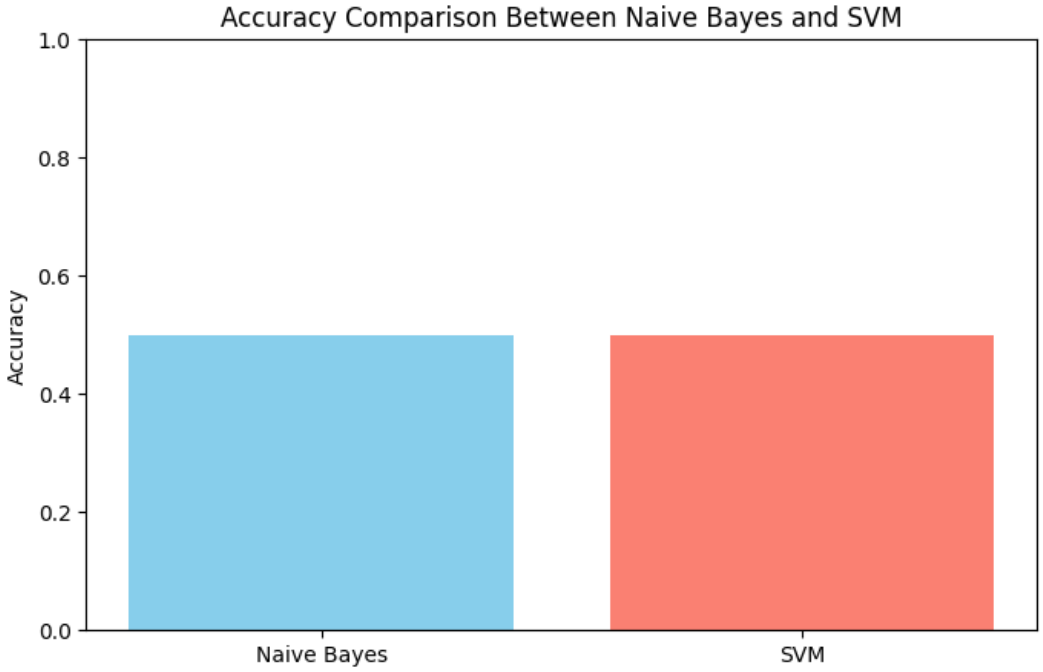
Tahun Ajaran	Jumlah Pendaftar Awal	Jumlah Pendaftar Tetap	Fakultas favorit
2017	680	651	FEB
2018	1078	1051	FEB
2019	1308	1274	FEB
2020	1292	1227	FEB
2021	1342	1270	PGSD
2022	1626	1513	PGSD
2023	2005	1917	PGSD
2024	1682	1605	PGSD

Berikut adalah hasil perbandingan peramalan Naïve Bayes dan Support Vector Machine:

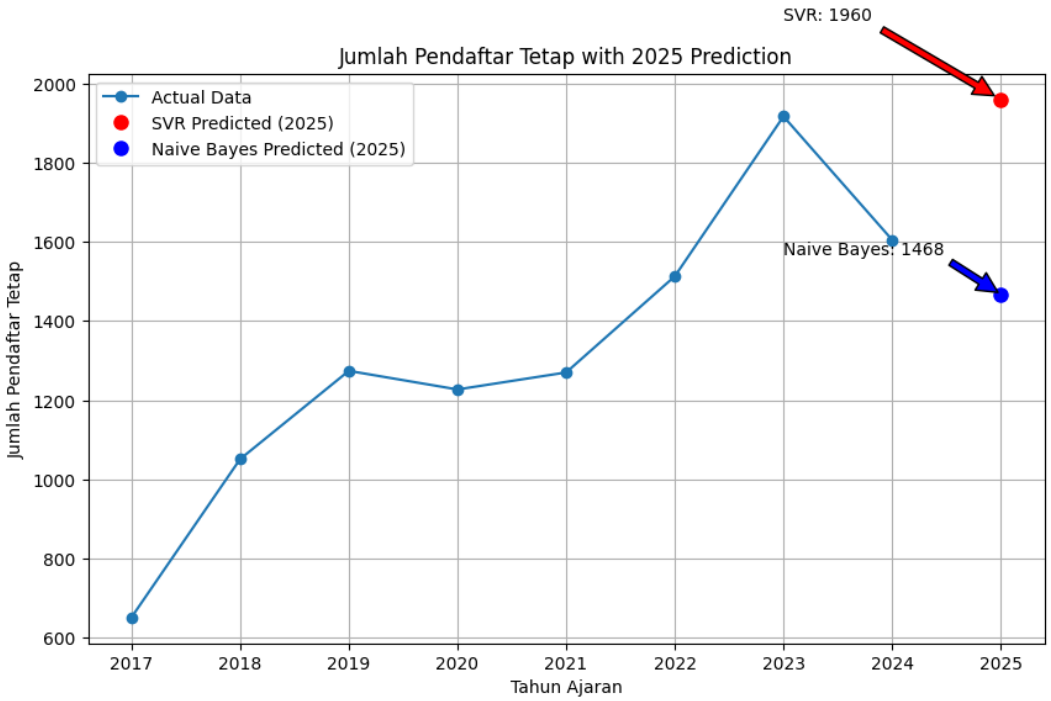
Tabel 2. Transformasi Data

Metode	**Prediksi Kategori**	**Prediksi Jumlah Pendaftar Tetap 2025**	**Akurasi**
Naïve Bayes	Tinggi	1542.5	50%
SVM (SVR)	-	1883	50%

Grafik batang dibuat untuk menampilkan perbandingan akurasi antara model Naïve Bayes dan SVM, dengan Naïve Bayes menghasilkan prediksi kategori dan SVM memberikan prediksi numerik. Perbandingan akurasi ini menunjukkan performa relatif dari masing-masing model pada data uji.



Grafik tren jumlah pendaftar tetap dibuat berdasarkan data historis, dengan prediksi tahun 2025 ditampilkan sebagai titik tambahan pada grafik. Nilai prediksi dari Naïve Bayes dan SVM memberikan perspektif tentang bagaimana tren kemungkinan berlanjut di masa mendatang.



KESIMPULAN

Kesimpulan penelitian ini menunjukkan bahwa kedua metode, Naïve Bayes dan Support Vector Machine, dapat diunggulkan untuk memprediksi jumlah pendaftar tetap. Model Naïve Bayes, dengan tingkat akurasi moderat sebesar 50%, mampu melakukan prediksi dalam bentuk kategori dengan estimasi pendaftar tetap pada tahun 2025 dalam kategori Tinggi, yaitu sekitar 1542.5. Di sisi lain, SVM dalam bentuk Support Vector Regression (SVR) juga mencapai akurasi sebesar 50% namun memberikan prediksi numerik yang lebih rinci dengan estimasi sebesar 1883 pendaftar. Metode SVM cocok digunakan pada data dengan tren peningkatan dan lebih tepat untuk prediksi numerik. Perbandingan kedua metode ini menegaskan bahwa Naïve Bayes efektif untuk prediksi kategori sederhana, sementara SVM lebih unggul untuk prediksi numerik presisi. Hasil prediksi ini sangat penting bagi institusi pendidikan dalam perencanaan kapasitas penerimaan dan pengelolaan sumber daya di tahun mendatang.

REFERENSI

[1] A. Nikolay, G. Anindya, and G. I. Panagiotis, “Deriving the pricing power of product features by mining consumer reviews,” Manage. Sci., vol. 57, no. 8, pp. 1485–1509, 2011, doi: 10.1287/mnsc.1110.1370.

[2] M. L. Ashari and M. Sadikin, “Prediksi Data Transaksi Penjualan Time Series Menggunakan Regresi Lstm,” J. Nas. Pendidik. Tek. Inform., vol. 9, no. 1, p. 1, 2020, doi: 10.23887/janapati.v9i1.19140.

[3] A. Benlahbib and E. H. Nfaoui, “An Unsupervised Approach for Reputation Generation,” in Procedia Computer Science, 2019, vol. 148, pp. 80–86. doi: 10.1016/j.procs.2019.01.011.

[4] W. M. Wang, Z. G. Tian, Z. Li, J. W. Wang, A. Vatankhah Barenji, and M. N. Cheng, “Supporting the construction of affective product taxonomies from online customer reviews: an affective-semantic approach,” J. Eng. Des., vol. 30, no. 10–12, pp. 445–476, 2019, doi: 10.1080/09544828.2019.1642460.

[5] G. A. Ruz, P. A. Henríquez, and A. Mascareño, “Sentiment analysis of Twitter data during critical events through Bayesian networks classifiers,” Futur. Gener. Comput. Syst., vol. 106, pp. 92–104, 2020, doi: 10.1016/j.future.2020.01.005.

[6] R. Alfrjani, T. Osman, and G. Cosma, “A Hybrid Semantic Knowledgebase-Machine Learning Approach for Opinion Mining,” Data Knowl. Eng., vol. 121, pp. 88–108, 2019, doi: 10.1016/j.datak.2019.05.002.

[7] G. W. N. Wibowo and M. A. Manan, “Penerapan Algoritma Naive Bayes Untuk Prediksi Heregistrasi Calon Mahasiswa Baru,” JTINFO J. Tek. …, vol. 1, no. 1, pp. 1–10, 2022, [Online]. Available: https://journal.unisnu.ac.id/JTINFO/article/view/126

[8] B. D. Prasetya, F. S. Pamungkas, and I. Kharisudin, “Pemodelan dan Peramalan Data Saham dengan Analisis Time Series menggunakan Python,” Prism. Pros. Semin. Nas. Mat., vol. 3, pp. 714–718, 2020, [Online]. Available: https://journal.unnes.ac.id/sju/index.php/prisma/article/view/38116

[9] H. Himawan and P. D. P. Silitonga, "Comparison of forecasting accuracy rate of exponential smoothing method on admission of new students," J. Crit. Rev., vol. 7, no. 2, pp. 268–274, 2020, doi: 10.31838/jcr.07.02.50.
[10] S. J. Taylor and B. Letham, "Business Time Series Forecasting at Scale," PeerJ Prepr. 5e3190v2, vol. 35, no. 8, pp. 48–90, 2017.
[11] M. S. R. Maulana, "IMPLEMENTASI LITERASI DIGITAL DI SEKOLAH, SEBUAH KENISCAYAAN," Ekp, 2017. https://disdikkbb.org/news/implementasi-literasi-digital-di-sekolah-sebuah-keniscayaan/ (accessed Aug. 11, 2021).
[12] M. Al-Smadi, O. Qawasmeh, M. Al-Ayyoub, Y. Jararweh, and B. Gupta, "Deep Recurrent neural network vs. support vector machine for aspect-based sentiment analysis of Arabic hotels' reviews," J. Comput. Sci., vol. 27, pp. 386–393, 2018, doi: 10.1016/j.jocs.2017.11.006.
[13] J. W. G. Putra, "Pengenalan Konsep Pembelajaran Mesin dan Deep Learning," Comput. Linguist. Nat. Lang. Process. Lab., vol. 4, pp. 1–235, 2019, [Online]. Available: https://www.researchgate.net/publication/323700644
[14] K. K. Aggarwal, Y. Singh, P. Chandra, and M. Puri, "Bayesian Regularization in a Neural Network Model to Estimate Lines of Code Using Function Points," J. Comput. Sci., vol. 1, no. 4, pp. 505–509, 2005, doi: 10.3844/jcssp.2005.505.509.
[15] A. Yadav, C. K. Jha, A. Sharan, and V. Vaish, "Sentiment analysis of financial news using unsupervised approach," Procedia Comput. Sci., vol. 167, no. 2019, pp. 589–598, 2020, doi: 10.1016/j.procs.2020.03.325.
[16] J. Jabbar, I. Urooj, W. Junsheng, and N. Azeem, "Real-time sentiment analysis on E-Commerce application," in Proceedings of the 2019 IEEE 16th International Conference on Networking, Sensing and Control, ICNSC 2019, 2019, pp. 391–396. doi: 10.1109/ICNSC.2019.8743331.
[17] N. Gali, R. Mariescu-Istodor, D. Hostettler, and P. Fränti, "Framework for syntactic string similarity measures," Expert Syst. Appl., vol. 129, pp. 169–185, 2019, doi: 10.1016/j.eswa.2019.03.048.
[18] F. Nurhuda, S. Widya Sihwi, and A. Doewes, "Analisis Sentimen Masyarakat terhadap Calon Presiden Indonesia 2014 berdasarkan Opini dari Twitter Menggunakan Metode Naive Bayes Classifier," J. Teknol. Inf. ITSmart, vol. 2, no. 2, p. 35, 2016, doi: 10.20961/its.v2i2.630.