INNOVATIVE: Journal Of Social Science Research
Volume 4 Nomor 1 Tahun 2024 Page 7537-7548
E-ISSN 2807-4238 and P-ISSN 2807-4246
Website: https://j-innovative.org/index.php/Innovative

Optimasi Feature Selection Text Mining: Stemming dan Stopword untuk Sentimen Analisis Aplikasi SatuSehat

Diky Wardhani[1], Rika Astuti [2✉], Dedi Dwi Saputra [3]
Program Studi Teknologi Informasi, Universitas Siber
IndonesiaEmail: rika.astuti@cyber-univ.ac.id[2✉]

Abstrak

Aplikasi SatuSehat merupakan hasil pengembangan dan transformasi dari Peduli Lindungi yang dilakukan oleh KEMENKES (kementerian kesehatan) dengan tujuan untuk mencatat data kesehatan masyarakat. Dengan berubahnya aplikasi Peduli Lindungi menjadi SatuSehat, beragam ulasan di layangkan oleh pengguna aplikasi ini di PlayStore. Penelitian ini dilakukan untuk membandingkan dan mengoptimasikan Feature Selection pada Text Mining untuk mendapatkan hasil yang paling optimal dari kedua fitur tersebut. 2000 data set didapatkan dengan metode Scrapping pada ulasan PlayStore. Lalu data tersebut diterapkan metode SMOTE dan Pre-Processing dengan Feature Selection, Stemming dan Stopword sehingga kedua fitur dapat dibandingkan dan dicari hasil yang optimal. Hasil penelitian ini maka bisa diperoleh hasil saat menggunakan Feature Selection steaming hasilnya akurasinya mendapatkan 93,43% dan presisi mendapatkan 88,42% sedangkan saat menggunakan feature selection stopword hasil yang didapatkan adalah nilai akurasinya mendapatkan 89,19% dan presisi mendapatkan 82,23%, dan jika menggunkan stopword dan stemming dilakukan secara bersamaan maka hasilnya nilai akurasinya mendapatkan 92,56% dan presisi mendapatkan 95,46%. Dan hasil teroptimal diperoleh paling optimal saat menggunakan stemming dan stopword digunakan secara bersamaan.

Kata Kunci : Analisis Sentimen, Naïve Bayes, SMOTE, Stemming, Stopword

Abstract

The SatuSehat application is the result of development and transformation from Peduli Lindungi conducted by the Ministry of Health (KEMENKES) with the aim of recording public health data. With the transformation of the Peduli Lindungi application into SatuSehat, various reviews have been provided by users of this application on the PlayStore. This research was conducted to compare and optimize Feature Selection in Text Mining to obtain the most optimal results from both features. A total of 2000 datasets were obtained using the Scrapping method on PlayStore reviews. Then, the data was applied with the SMOTE method and Pre-Processing using Feature Selection, Stemming, and Stopword, so that both features could be compared and the optimal results could be found. The results of this research show that when using Feature Selection with stemming, the accuracy obtained was 93.43%, and the precision obtained was 88.42%. On the other hand, when using Feature Selection with stopword, the accuracy obtained was 89.19%, and the precision obtained was 82.23%. Finally, when both stopword and stemming were used together, the accuracy obtained was 92.56%, and the precision obtained was 95.46%. The most optimal results were obtained when stemming and stopword were used simultaneously.

Keywords: Sentiment Analysis, Naïve Bayes, SMOTE, Stemming, Stopword

PENDAHULUAN

Pada era globalisasi dan perkembangan teknologi informasi yang pesat, penelitian dalam berbagai bidang ilmu menjadi semakin penting. Salah satu bidang yang terus mengalami perkembangan dan perubahan adalah Data Mining. Dalam konteks ini, penelitian ini bertujuan untuk menganalisis dan mengeksplorasi komparasi dan optimasi feature selection text mining: stemming dan stopword untuk sentimen analisis aplikasi SatuSehat. Satusehat merupakan aplikasi Kesehatan masyarakat dan hasil transformasi serta pengembangan dari aplikasi peduli lindungi yang dilakukan oleh kementrian Kesehatan (kemenkes) dengan tujuan untuk mencatat dan memonitor kondisi kesehatan diri maupun orang-rang terdekat(Rokom, 2023). Namun dengan bertransformasinya Pedulilindungi menjadi Satusehat banyak pengguna yang melayangkan ulasannya kepada Satusehat di Playstore.

Sentiment Analysis (Analisis Sentimen) merupakan suatu cabang dari Natural Languange Processing (NLP) yang membangun suatu sistem untuk mengidentifikasi dan mengekstrak opini dalam bentuk teks(adminlp2m, 2022). Analisa Sentimen membantu membuat gambaran respon masyarakat dengan mengkategorikan ulasan menjadi komplen dan bukan komplen sehingga hasil opini yang diperoleh nantinya dapat dievaluasi oleh aplikasi Satusehat (Pratama Putra et al., 2022).

Text mining merupakan sebuah teknologi untuk mengelola data teks dengan tujuan mendapatkan informasi secara otomatis. Dengan menggunakan text mining, kita dapat menghasilkan informasi baru melalui analisis data teks yang bersifat semi terstruktur atau tidak terstruktur, biasanya dalam jumlah yang besar.

Teknologi ini sangat bermanfaat bagi pekerjaan manusia mengingat jumlah data teks dan dokumen yang ada di aplikasi web, aplikasi digital, dan media sosial yang semakin meningkat. Data-data tersebut sering kali bersifat besar dan kurang terstruktur, sehingga membutuhkan waktu yang lama untuk menganalisis informasi di dalamnya.

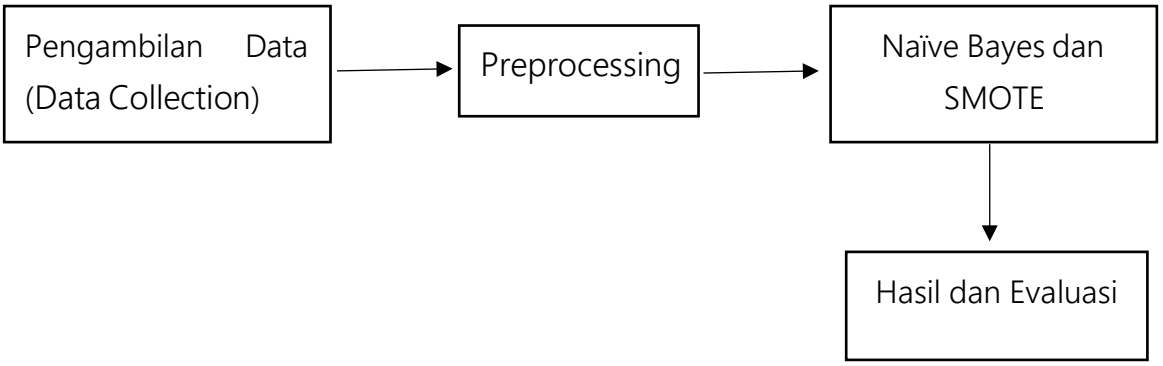
Naive Bayes mwerupakan salah satu metode klasifikasi supervised yang paling umum yang dapat digunakan untuk klasifikasi data teks. Sebelum melakukan itu, pertama-tama kita perlu melihat apa itu vektor fitur. Untuk mengklasifikasikan, pertama-tama kita perlu memilih fitur dari data. Dalam klasifikasi teks, vektor fitur juga dikenal sebagai vektor konsep dan merupakan struktur utama dari proses pelatihan dan klasifikasi. Semua teks ulasan diubah menjadi vektor konsep yang diproses oleh pengklasifikasi. Biasanya, vektor istilah dibangun berdasarkan kosa kata unik yang dibangun dari kumpulan data pelatihan, tanpa kata duplikat dalam kosa kata. Ukuran vektor konsep sesuai dengan ukuran kosakata (ZHAOYU Yi Shang, 2014).

Karena algoritme naive bayes tidak memperhitungkan ketidakseimbangan data, hal ini berdampak pada kurangnya efisiensi prediktif dan meningkatkan bias pada kelas dengan data lebih banyak (kelas mayoritas).

SMOTE merupakah teknik oversampling yang bertugas dengan menaikkan total instance kelas positif melalui replikasi data secara acak, menyeimbangkan total data positif dengan data negatif. Untuk memakai data sintetik, replikasi data dilakukan dalam kelas yang lebih sedikit. Algoritma SMOTE betugas mencari k tetangga sekitar untuk kelas positif dan lalu membuat duplikat sintetik dari data antara k kelas yang diambil secara sembarang dan kelas positif dengan persentase yang diinginkan. Cara ini mungkin bisa mengatasi masalah ketidakseimbangan data (Suryana & Tri Prasetio, 2020).

METODE PENELITIAN

Metode yang digunakan untuk menganalisis masalah adalah sebagai berikut:



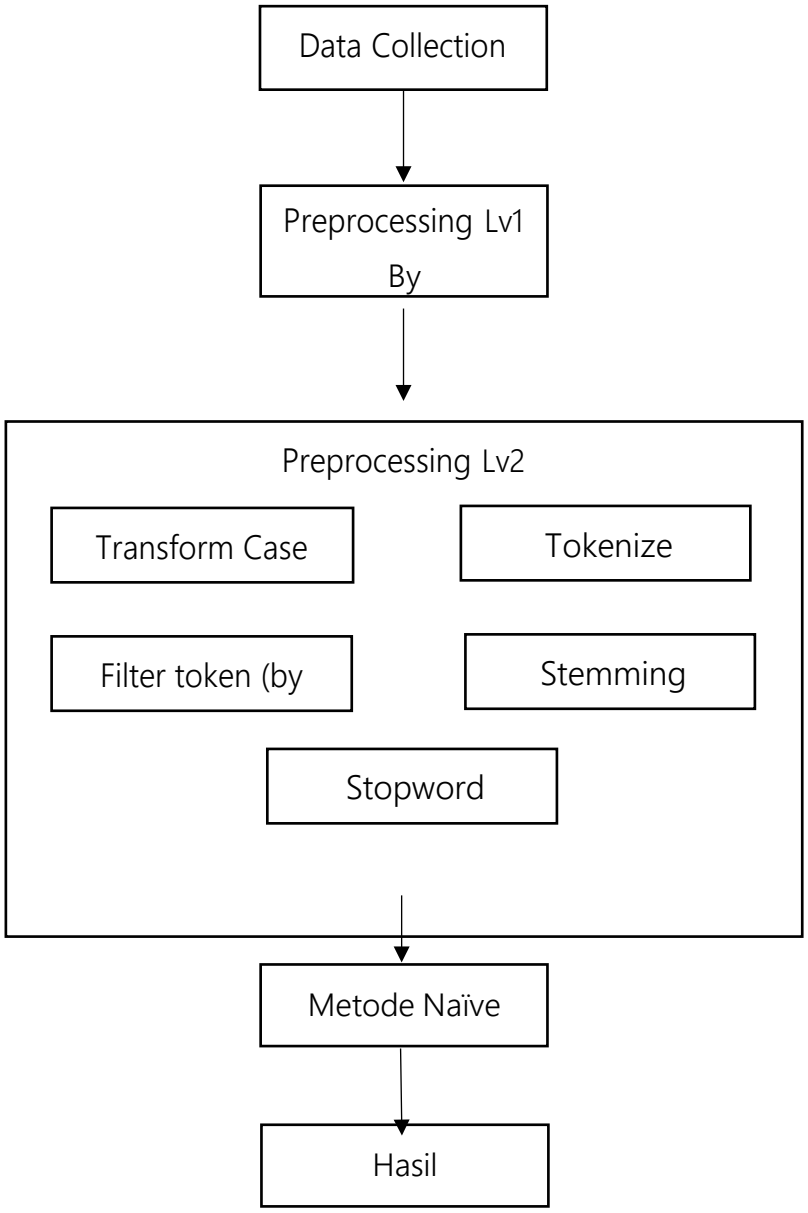
Sumber:Penelitian

Gambar 1.Proses Penelitian

1. Pengambilan Data (Data Collection), Penulis menggunakan metode Scrapping, data yang digunakan berformat file excel, dan isi data tersebut mengenai ulasan aplikasi Satusehat yang diperoleh dari playstore dengan kategori komplain dan bukan komplain melalui metode scrapping memakai Python. Pengambilan dataset dengan metode scrapping dilakukan sebanyak 2000 data ulasan aplikasi Satusehat di Playstore dengan melalui filterisasi sehingga data yang diperoleh merupakan data yang relevan dan terbaru pada saat data tersebut diambil.
2. Data pre-processing adalah teknik yang melibatkan transformasi data mentah menjadi format yang mudah dimengerti. Proses data pre-processing dibutuhkan untuk mengatasi berbagai permasalahan, contoh data yang tidak akurat, redudansi datan dan nilai data yang hilang(Kurniawan Rachmat dkk., 2023),(Kotsiantis dkk., 2014). Pada tahap ini data dilakukan preprocessing sebanyak 2 tahap. Tahap pertama, data dipreprocessing melalui gataframework dengan menghapus regular expression, stemming, dan stopword sesuai kaidah Bahasa Indonesia. Tahap dua preprocessing, Langkah ini sering digunakan untuk mengubah data teks menjadi sentimen. Pada tahap ini, ada berberapa teknik filtering yang bisa digunakan, misalnya:
    a. Transform case: Semua kata dan karakter(abjad) kapital dalam data diubah sebagai karakter(abjed)kecil.
    b. Tokenize: Metode untuk membentuk beberapa token melalui pemisahan data teks.
    c. Filter token (by lenght): Memfilter token berdasarkan panjangnya karakter.
    d. Stemming: proses menghilangkan imbuhan dan menerjemahkan kata sesuai objek sentimen.

e. Stopword: proses mengeliminasi kata yang kurang sesuai dengan objek sentiment.

3. Metode yang dipakai dalam penelitian ini adalah klarifikasi Naive Bayes. (Deepa et al., 2022), dimana pada tahap ini bahan dianalisis kemudian diterapkan model sesuai tipe datanya.



Sumber: Penelitian (2023)

Gambar 2. Alur Proses

HASIL DAN PEMBAHASAN

1. Analisa Pengujian

Sistem yang telah dikembangkan merupakan sistem yang dapat digunakan untuk menganalisis sentimen dari ulasan yang merupakan keluhan (complain) dan bukan keluhan (non-complain) pada aplikasi Satusehat. Sistem ini bekerja dengan melakukan pengambilan data ulasan dari Play Store dan kemudian memprosesnya dalam sistem. Perhitungan dilakukan dengan menggunakan langkah-langkah yang dijelaskan dalam metode penelitian ini, diantaranya:

a. Pengambilan Data (Data Collection)

Data tersebut merupakan data ulasan yang terdapat pada Playstore dan diambil dengan proses scraping memakai Python dengan memanfaatkan package googleplay scraper dan dapatkan sebanyak 2000 dataset.

Tabel1.Proses Scrapping

No	Proses	Deskripsi
1	Install Package Google Play Scraper dan Pandas	Mengistal package yang diperlukan untuk scraping pada ulasan playstore
2	Search ID Satusehat	Mengambil ID aplikasi yang akan digunakan sebagai data set ulasan pada playstore
3	Set Attribut Ulasan Pada ID Tujuan	Memfilter dataset ulasan yang akan di scraping dengan filter terbaru, relevan, score, dan bahasa indonesia
4	Write data to CSV	Mengkonversi dataset menjadi format file dalam bentuk excel atau CSV

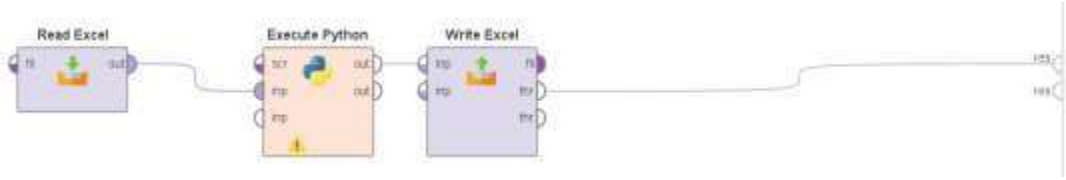
Sumber : Penelitan (2023)

b. Preprocessing

Data pre-processing adalah teknik data mining yang melibatkan transformasi data mentah menjadi format yang mudah dipahami. Proses data pre-processing dipergunakan untuk mengatasi berbagai permasalahan, seperti data yang tidak akurat, redudansi data dan nilai data yang hilang(Kurniawan Rachmat dkk., 2023), (Kotsiantis dkk., 2014). Di tahap ini setelah data teks diolah menjadi sentiment analisis oleh tahap preprocessing, metode SMOTE upsampling digunakan untuk menangani ketidak seimbangan (imbalance) data, karena algoritma naïve bayes tidak melihat imbalanceing data sehingga berdampak terhadap kurangnya peforma prediksi dan akan terjadi peningkatan bias pada kelas yang datanya lebih banyak. Adapun prosesnya sebagai

berikut:

Gambar 3. Preprocessing By Gataframework



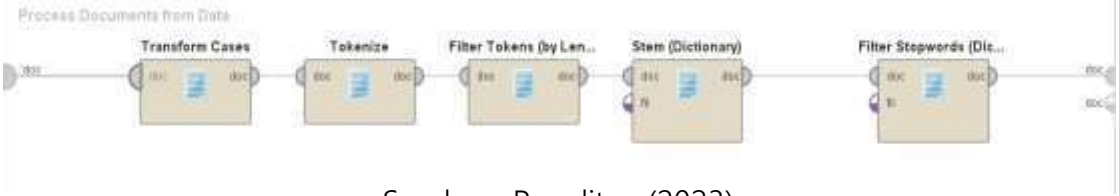
Sumber : Penelitan (2023)

Tabel 2. Alur Preprocessing Lv2 Analsis Sentimen

No	Proses	Deskripsi
1	Read Excel	Membaca file tipe excel dataset yang sudah di preprocessing oleh Gataframework sesuai kaidah bahasa Indonesia
2	Select attribut	Memilih dan memfilter satu attribut pada file dataset yaitu text
3	Set Role	Mengubah peran dari beberapa attribut yang ada pada dataset
4	Proses Document	Mengubah dataset menjadi dokumen dan melakukan beberapa filter contohnya: Filter Token (by length, Stemming, Stopword), Transform case, Tokenize,
5	SMOTE Upsampling	Menangani ketidak seimbangan data
6	Cross Validation	Menganalisis dokumen dan menerapkan klasifikasi algoritma naïve bayes

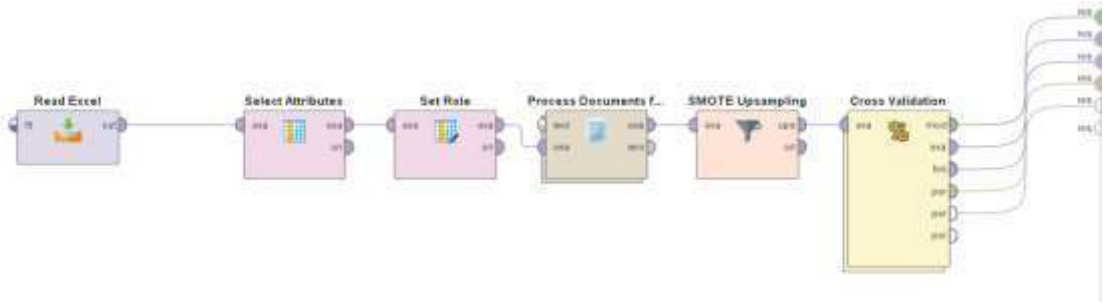
Sumber : Penelitan (2023)

Gambar 4. Filtering Proses Dokumen



Sumber : Penelitan (2023)

Gambar 5. Preprocessing Lv2 dengan SMOTE



Sumber : Penelitan (2023)

Dalam tahap ini, dataset telah dikumpulkan dan diproses dengan filtering sehingga hanya kata-kata tanpa tanda baca yang tersisa didalam data.

- Transform Case: Semua kata dan karakter(abjad) kapital dalam data diubah sebagai karakter(abjad) kecil.
- Tokenize: Metode untuk membentuk beberapa token melalui pemisahan data teks.
- Filter Token (by Lenght): Memfilter token berdasarkan panjanganya yaitu minimal 3 karakter dan maksimal 24 karakter.
- Stemming: Menghilangkan imbuhan dan menejermahkan kata sesuai dengan kebutuhan objek.
- Stopword: Menghapus kata yang tidak dibutuhkan sesuai dengan objek.

c. Naïve Bayes

Untuk proses menghitung akurasi menggunakan algoritma Naive Bayes melibatkan langkah-langkah seperti pelatihan (training) dan pengujian (testing). Langkah pertama adalah memberikan mesin sebuah skema data. Mesin akan belajar dari pola data ini dan menggunakan hasilnya sebagai referensi untuk memodelkan pola baru dengan akurasi tinggi.

Gambar 6. Preprocessing feature selection stemming

Sumber : Penelitan (2023)

d. Analisa Hasil Sentimen

e. Gambar 7. Preprocessing feature selection stemming

Sumber : Penelitan (2023)

Gambar 8. Preprocessing feature selection stemming

accuracy: 93.43% +/- 1.10% (micro average: 93.43%)

	true Complaint	true Not Complaint	class precision
pred. Complaint	1692	0	100.00%
pred. Not Complaint	256	1948	88.38%
class recall	86.86%	100.00%	

Sumber : Penelitan (2023)

Gambar 9. Preprocessing feature selection stopword

precision: 88.42% +/- 1.72% (micro average: 88.38%) (positive class: Not Complaint)

	true Complaint	true Not Complaint	class precision
pred. Complaint	1692	0	100.00%
pred. Not Complaint	256	1948	88.38%
class recall	86.86%	100.00%	

Sumber : Penelitan (2023)

Hasil penggunaan feature selection stemming pada metode smote dan Naïve Bayes, nilai akurasinya mendapatkan 93,43% dan presisi mendapatkan 88,42%

Gambar 10. Preprocessing feature selection stopword

Sumber : Penelitan (2023)

Gambar 11. Preprocessing feature selection stopword

accuracy: 89.19% +/- 1.31% (micro average: 89.19%)

	true Complaint	true Not Complaint	class precision
pred. Complaint	1527	0	100.00%
pred. Not Complaint	421	1948	82.23%
class recall	78.39%	100.00%	

Sumber : Penelitan (2023)

Gambar 11. Preprocessing feature selection stemming dan stopword

precision: 82.26% +/- 1.80% (micro average: 82.23%) (positive class: Not Complaint)

	true Complaint	true Not Complaint	class precision
pred. Complaint	1527	0	100.00%
pred. Not Complaint	421	1948	82.23%
class recall	78.39%	100.00%	

Sumber : Penelitan (2023)

Hasil penggunaan feature selection stopword pada metode smote dan Naïve Bayes, nilai akurasinya mendapatkan 89,19% dan presisi mendapatkan 82,23%

Gambar 12. Preprocessing feature selection stemming dan stopword

Sumber : Penelitan (2023)

Gambar 12. Preprocessing feature selection stemming dan stopword

accuracy: 92.56% +/- 1.53% (micro average: 92.56%)

	true Complaint	true Not Complaint	class precision
pred. Complaint	1864	206	90.05%
pred. Not Complaint	84	1742	95.40%
class recall	95.69%	89.43%	

Sumber : Penelitan (2023)

Gambar 13. Performance Precision

precision: 95.46% +/- 2.28% (micro average: 95.40%) (positive class: Not Complaint)

	true Complaint	true Not Complaint	class precision
pred. Complaint	1864	206	90.05%
pred. Not Complaint	84	1742	95.40%
class recall	95.69%	89.43%	

Sumber : Penelitan (2023)

Hasil penggunaan feature selection stemming dan stopword pada metode smote dan Naïve Bayes, nilai akurasinya mendapatkan 92,56% dan presisi mendapatkan 95,46%

SIMPULAN

Dari hasil penelitian ini maka bisa dperoleh hasil saat menggunakan feature selection steaming hasilnya akurasinya mendapatkan 93,43% dan presisi mendapatkan 88,42% sedangkan saat menggunakan feature selection stopword hasil yang didapatkan adalah nilai akurasinya mendapatkan 89,19% dan presisi mendapatkan 82,23%, dan jika menggunkan stopword dan stemming dilakukan secara bersamaan maka hasilnya nilai akurasinya mendapatkan 92,56% dan presisi mendapatkan 95,46%. Dari kedua perbandingan fitur stemming dan stopword

hasil terbesar diperoleh ketika menggunakan fitur stemming, namun hasil teroptimal untuk dataset yang ada yaitu diperoleh ketika menggunakan kedua feature selection tersebut secara bersamaan.

DAFTAR PUSTAKA

adminlp2m. (2022, Februari 21). ANALISIS SENTIMEN (SENTIMENT ANALYSIS): DEFINISI, TIPE DAN CARA KERJANYA. lp2m.uma.ac.id. https://lp2m.uma.ac.id/2022/02/21/analisis-sentimen-sentiment-analysis-definisi-tipe-dan-cara-kerjanya/

Cahyani, L. (2023). APLIKASI TEXT MINING DI BIDANG PENDIDIKAN (F. Andriansyah, Ed.). CV Literasi Nusantara Abadi. https://books.google.co.id/books?id=UjmqEAAAQBAJ

Deepa, N., Sathya Priya, J., & Devi, T. (2022). Towards applying internet of things and machine learning for the risk prediction of COVID-19 in pandemic situation using Naive Bayes classifier for improving accuracy. Materials Today: Proceedings, 62, 4795–4799. https://doi.org/10.1016/j.matpr.2022.03.345

Hendra, A. (2021). Analisis Sentimen Review Halodoc Menggunakan Nai ¨ ve Bayes Classifier.

Dalam JISKa (Vol. 6, Nomor 2). MEI.

Suryana, N., & Tri Prasetio, R. (2020). IJCIT (Indonesian Journal on Computer and Information Technology) Penanganan Ketidakseimbangan Data pada Prediksi Customer Churn Menggunakan Kombinasi SMOTE dan Boosting. Dalam IJCIT (Indonesian Journal on Computer and Information Technology) (Vol. 6, Nomor 1). https://creativecommons.org/licenses/by-sa/4.0/

Kotsiantis, S., Kanellopoulos, D., & Pintelas, P. E. (2014). Data Preprocessing for Supervised Learning A Big Data Scale Analysis Framework to Support Customized and Personalized Learning Environments View project Recent Transport Protocols in Wired Networks and Internet View project. https://www.researchgate.net/publication/228084519

Kurniawan Rachmat, B., Suwarisman, A., Afriyanti, I., Wahyudi, A., & Saputra, D. D. (2023).) 2023 1,2,3,4,5 Program Studi Sistem Informasi. Jurnal Teknologi Informasi dan Komunikasi), 7(1). https://doi.org/10.35870/jti

Pratama Putra, A., Pratama, Y., Kharisma Krisnadi, E., Purnamasari, I., & Dwi Saputra, D. (2022). Text Mining untuk Sentimen Analisis dengan Metode Naïve Bayes, SMOTE, N-Gram dan AdaBoost Pada Twitter CommuterLine. Dalam Jurnal Sains Komputer & Informatika (J-SAKTI (Vol. 6, Nomor 2). https://doi.org/http://dx.doi.org/10.30645/j-sakti.v6i2.506

Rokom. (2023, Maret 10). Besok PeduliLindungi Resmi Bertransformasi Menjadi SATUSEHAT Mobile. sehatnegeriku.kemkes.go.id. https://sehatnegeriku.kemkes.go.id/baca/rilis-media/20230228/2042474/besok-pedulilindungi-resmi-bertransformasi-menjadi-satusehat-mobile/

ZHAOYU Yi Shang, B. L. (2014). NAIVE BAYES ALGORITHM FOR TWITTER SENTIMENT ANALYSIS AND ITS IMPLEMENTATION IN MAPREDUCE.