INNOVATIVE: Journal Of Social Science Research
Volume 5 Nomor 1 Tahun 2025 Page 2338-2347
E-ISSN 2807-4238 and P-ISSN 2807-4246
Website: https://j-innovative.org/index.php/Innovative

Klasifikasi Sentimen Postingan Sosial Media Menggunakan Machine Learning Random Forest dan Naïve Bayes

Tommy Dwi Putra[1✉], Dewi Oktafiani[2]
STMIK AMIKOM Surakarta
Email: tommy@dosen.amikomsolo.ac.id[1✉]

Abstrak

Penggunaan media sosial di Indonesia berkembang pesat sehingga menimbulkan dampak positif seperti peningkatan kreativitas dan kemudahan berkomunikasi, serta dampak negatif seperti kecanduan dan kesepian. Penelitian ini berfokus pada klasifikasi sentimen postingan media sosial menggunakan algoritma Random Forest dan Naïve Bayes. Tujuannya adalah untuk memahami sentimen di media sosial, yang berguna bagi organisasi ketika merespons opini publik dan mengambil keputusan strategis. Penelitian ini membandingkan kinerja kedua algoritma tersebut menggunakan dataset besar dari Twitter. Hasil penelitian menunjukkan bahwa algoritma Naïve Bayes mencapai akurasi sebesar 90,41%, sedangkan algoritma Random Forest mencapai akurasi sebesar 39.74%.
Kata Kunci : *Klasifikasi, Sosial Media, Random Forest, Naïve Bayes*

Abstract

The use of social media in Indonesia is growing rapidly, causing positive impacts such as increased creativity and ease of communication, as well as negative impacts such as addiction and loneliness. The study focused on the sentiment classification of social media posts using the Random Forest and Naïve Bayes algorithms. The goal is to understand sentiment on social media, which is useful for organizations when responding to public opinion and making strategic decisions. This study compared the performance of the two algorithms using a large dataset from Twitter. The results show that the Naïve Bayes algorithm achieves an accuracy of 90.41%, while the Random Forest algorithm achieves an accuracy of 39.74%.
Keyword: *Classification, Social Media, Random Forest, Naïve Bayes*

PENDAHULUAN

Berkaitan dengan peningkatan signifikan dalam penggunaan media sosial dalam komunikasi sehari-hari. Pada Januari 2021, Indonesia termasuk dalam 10 negara teratas dalam tingkat kecanduan media sosial ke-9. Sekitar 170 juta orang aktif menggunakan Internet dan media sosial, dan masyarakat Indonesia menghabiskan 8 jam 52 menit sehari. Pengguna menggunakan media sosial untuk memperkenalkan diri, berkomunikasi, berkolaborasi, berbagi informasi atau berinteraksi dengan pengguna lain.

Dampak positif dari internet dan media sosial memberikan berbagai dampak, seperti peningkatan kreativitas individu dalam membuat konten media dan kemampuan mengirim dan menerima pesan dengan berbagai pihak kapan saja dan dimana saja. Namun tidak terelakkan bahwa kemunculan media sosial dan penggunaan internet yang berlebihan memberikan dampak negatif seperti kecanduan, kesepian, dan berkurangnya kemungkinan berinteraksi dengan orang lain.

Analisis sentimen adalah proses komputer yang secara otomatis memanipulasi, membedah, dan mencerna data teks untuk memperoleh data opini yang terkandung dalam kalimat opini atau komentar, kumpulan sikap dan perasaan manusia (untuk mendeskripsikan seseorang, peristiwa, atau topik), yang menjadi fokus banyak industri dari layanan analisis sentimen. Sentimen dalam postingan media sosial seringkali ambigu dan tidak terstruktur. Analisis sentimen manual menjadi tidak mungkin dilakukan secara efisien dengan jumlah data yang besar.

Posting berasal dari kata “post” yang berarti mengirim atau mempublikasikan. Namun saat ini istilah tersebut sering kita dengar di dunia maya ketika kita berselancar di media sosial. Emosi seseorang bisa muncul dari penggunaan media sosial. Naiknya emosi sangatlah berbeda dan naiknya emosi dapat menimbulkan dampak negatif. Walaupun emosi sangat kompleks, namun dapat dibagi menjadi beberapa kelompok seperti marah, sedih, takut, bahagia, terkejut, sedih.

Banyak percobaan dengan teknik data mining dilakukan dalam satu penelitian. Data mining adalah pembelajaran mesin, pengenalan pola, basis data, statistik, dan teknik visualisasi yang digunakan untuk memecahkan masalah segmentasi basis data yang besar. Pengelompokan sistem disebut klasifikasi, klasifikasi mempunyai dua tugas utama yaitu membuat model sebagai prototipe dalam memori dan menggunakannya untuk membuat rekomendasi/klasifikasi/prediksi berdasarkan objek data.

Random Forest adalah metode algoritmik yang digunakan untuk mengklasifikasikan kumpulan data. Metode partisi biner rekursif digunakan untuk sampai pada node terakhir

dari struktur pohon berdasarkan klasifikasi Random Forest (RF) dan pohon regresi. Algoritma Random Forest memiliki banyak keunggulan, seperti kemampuan menghasilkan kesalahan yang relatif rendah, kinerja klasifikasi yang sangat baik, kemampuan menangani data pelatihan yang besar secara efisien, dan merupakan cara yang efektif untuk memperkirakan data yang hilang. Naive Bayes merupakan pendekatan yang menggunakan teorema Bayes untuk menggabungkan informasi sebelumnya dengan informasi baru. Jadi ini merupakan algoritma klasifikasi yang sederhana namun memiliki akurasi yang tinggi.

Dua algoritma pembelajaran mesin, Random Forest dan Naive Bayes, digunakan untuk membandingkan kinerja media sosial setelah klasifikasi sentimen untuk membandingkan kinerja kedua algoritma dalam menyelesaikan masalah klasifikasi sentimen. Kinerja dibandingkan dengan mengukur metrik evaluasi untuk menentukan algoritma mana yang memberikan hasil terbaik. Perbandingan Random Forest dan Naive Bayes untuk mendapatkan pemahaman yang lebih lengkap tentang kinerja algoritma klasifikasi sentimen baru dan lama serta implikasi penggunaan algoritma ini dalam aplikasi praktis.

Di zaman yang semakin maraknya penggunaan media sosial, penting bagi kita untuk memahami emosi yang terkandung dalam setiap postingan. Dalam penelitian ini, penulis menggunakan pembelajaran mesin Random Forest dan Naïve Bayes untuk mengklasifikasikan sentimen media sosial Twitter. Berdasarkan uraian di atas maka penelitian ini fokus dilakukan dengan judul "Klasifikasi Sentimen Postingan Sosial Media Menggunkan Machine Learning Random Forest dan Naïve Bayes".

METODE PENELITIAN

Dataset

Datasetnya yang digunakan dalam penelitian ini adalah postingan sosial media dari berbagai Platform didapat dari web Kaggle. sebagian dari data ini digunakan dalam eksperimen yang dibagikan di Galeri Intelijen Cortana Microsoft. Kumpulan data ini disumbangkan oleh CrowdFlower, dengan 40.000 baris data tersedia untuk diunduh pada 15 Juli 2016. Penelitian membatasi dataset sebanyak 4000 baris data sebagai sampel dan hanya berfokus pada postingan/content.

Tabel 1. Detail Dataset

Tweet_id	Sentiment	author	Content
1956967341	Empty	xoshayzers	@tiffanylue i know i was listenin to bad habit...
1956967666	Sadness	wannamama	Layin n bed with a headache ughhhh...waitin on...
1956967696	Sadness	coolfunky	Funeral ceremony...gloomy friday...
1956967789	Enthusiasm	czareaquino	wants to hang out with friends SOON!
1956968416	Neutral	xkilljoyx	@dannycastillo We want to trade with someone...
1956968477	Worry	xxxPEACHESxxx	Re-pinging @ghostridah14: why didn't you go to...
1956968487	Sadness	ShansBee	I should be sleep, but im not! thinking about an old..
1956968636	Worry	mcsleazy	Hmmm. http://www.djhero.com/ is down
1956969035	Sadness	nic0lepaula	@charviray Charlene my love. I miss you
1956969172	Sadness	Ingenue_Em	@kelcouch I'm sorry at least it's Friday?

Pemrosesan Data

Tahap pemrosesan atau pengolahan data meliputi kegiatan membuat data sekaligus membersihkan data sehingga dapat digunakan untuk tahap modeling atau pemodelan data. Berikut tahapan pemrosesan data diantaranya:

1. Cleaning menghapus variabel yang tidak dibutuhkan, seperti URL, simbol.
2. Labeling memberi label positif, negatif dan netral.
3. Tokenizing memengal kalimat menjadi beberapa bagian atau kata.
4. Transform Case merubah semua huruf besar atau huruf kecil ataupun sebaliknya.
5. Stopword removal merupakan proses untuk menghilangkan kata yang tidak diperlukan.
6. Stemming menghilangkan imbuhan kata sehingga menjadi kata dasar.
7. Filter Tokens menghilangkan kata dengan panjang huruf tertentu.

Pembobotan Kata

Pembobotan kata atau Term Frequency- Inverse Document Frequecy (TF-IDF) adalah tahapan untuk memberikan nilai pada setiap kata yang ada bertujuan untuk menemukan frequency kemunculan term dan dikalikan dengan inverse dari frequency dokumen. Berikut ini rumus IDF pada persamaan berikut.

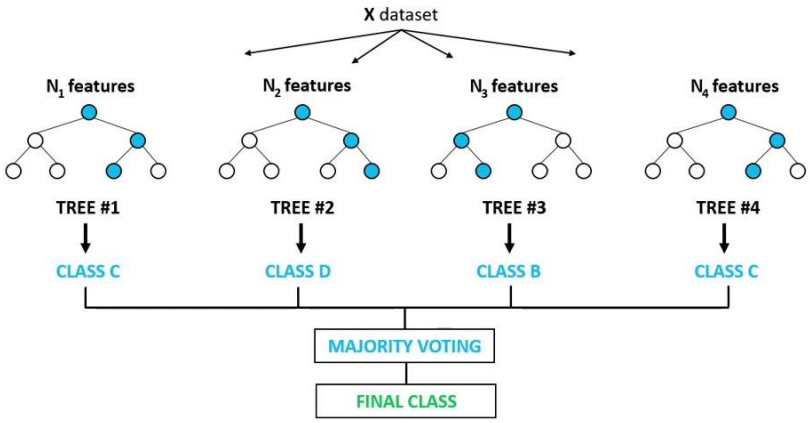
$$idf = log\left(\frac{n}{df_t}\right)$$

Dan rumus TF-IDF pada persamaan

$$tf - idf_{td}.idf_t$$

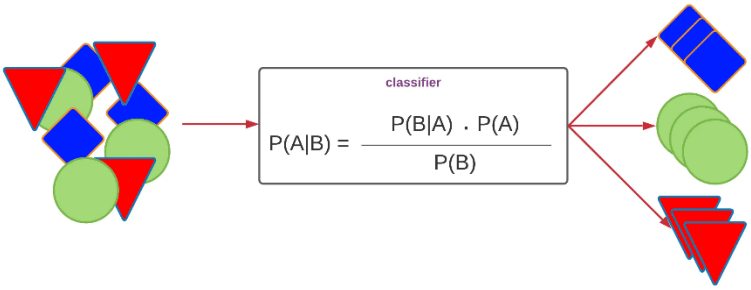
Klasifikasi

Pada tahap ini algoritme klasifikasi yang digunakan ialah Random Forest dan Naïve Bayes. Kemudian akan dibandingkan hasilnya. Random Forest (RF) merupakan salah satu jenis metode boostrap aggregating yang memiliki cara kerja dengan membangkitkan sejumlah tree dari data sample dimana pembuatan satu tree pada saat proses training tidak bergantung terhadap tree sebelumnya kemudian dalam pengambilan keputusannya diambil berdasarkan voting terbanyak. Cara kerja Random Forest dapat dilihat pada gambar di bawah ini.



Gambar 1. Ilustrasi Random Forest

Pengklasifikasi Naïve Bayes adalah pengklasifikasi pembelajaran yang diawasi karena memiliki supervisor (pengklasifikasian manual yang dilakukan manusia pada data yang digunakan dalam pelatihan) sebagai guru selama proses pembelajaran. Selain itu, kinerja Naïve Bayes memiliki waktu klasifikasi yang singkat, sehingga mempercepat proses sistem analisis sentiment. Cara kerja Naïve Bayes dapat dilihat pada gambar di bawah ini.



Gambar 2. Ilustrasi NB

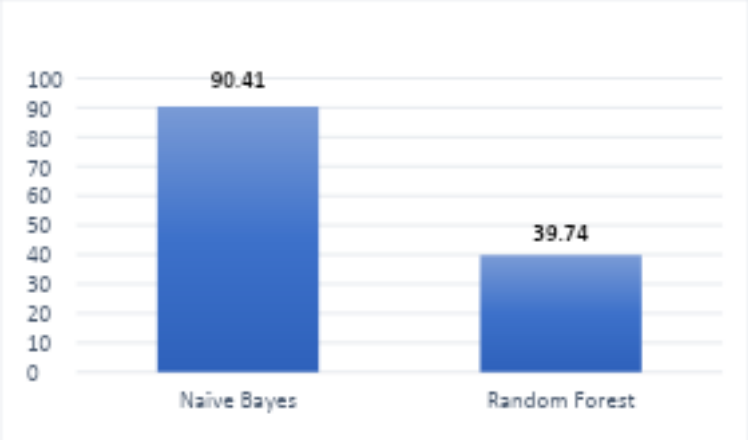
Evaluasi

Evaluasi untuk mengetahui nilai kegunaan dari model yang telah berhasil dibuat pada Langkah sebelumnya. Untuk tahap evaluasi akan menggunakan teknik *10 fold cross validation*. Perhitungan akurasi dilakukan sesuai dengan persamaan di bawah ini.

$$akurasi = \frac{\sum \ data\ uji\ benar}{\sum \ jumlah\ total\ data\ uji} 100$$

HASIL DAN PEMBAHASAN

Dengan menggunakan data sampel sebanyak 4000 baris yang mengandung 13 jenis sentiment. Naive Bayes mencatat akurasi sebesar 90,41%. Hal ini menegaskan pentingnya Naïve Bayes dalam menangani data berdimensi tinggi dan kompleks, seperti data media sosial, yang sering kali merupakan variabel tidak terstruktur dan terkait. Di sisi lain, Random Forest menunjukkan performa dengan akurasi 39,74%. Perbandingan dapat dilihat pada gambar di bawah ini.



Gambar 3. Grafik Perbandingan

Pengujian fungsional dilakukan dengan membagi dataset menjadi 70% data pelatihan dan 30% data pengujian. Evaluasi kinerja dilakukan menggunakan *10 fold cross validation* untuk menghitung tiga metrik utama akurasi. Naïve Bayes menunjukkan nilai yang sangat akurat dan mudah diingat, menunjukkan kemampuan algoritma ini dalam membuat prediksi yang akurat dan sensitif terhadap perubahan asumsi data. Performa untuk setiap algortima dapat dilihat tabel di bawah ini.

Tabel 2. Performa Vektor Naïve Bayes

Accuracy: 90.41% +/-0.43% (micro average: 90.41%)				
True	Netral	Negatif	Positif	Class Precision
Netral	9486	1420	1049	79.35%
Negatif	0	11913	80	99.33%
Positif	0	896	11093	92.53%
Class recall	100.00%	83.72%	90.76%	

Tabel 3. Performa Vector Random Forest

Accuracy: 39.74% +/-0.24% (micro average: 39.74%)

True	Netral	Negatif	Positif	Class Precision
Netral	0	0	0	0.00%
Negatif	1054	1581	1352	39.65%
Positif	0	0	6	100.00%
Class recall	0.00%	100.00%	0.44%	

Akurasi Naïve Bayes yang tinggi menunjukkan bahwa algoritma ini dapat mengurangi jumlah hasil yang baik, yaitu prediksi positif palsu. Hal ini sangat penting dalam konteks analisis emosional, dimana generalisasi adalah tentang kesalahpahaman perasaan atau pikiran. Daya ingat yang tinggi menunjukkan bahwa Naïve Bayes lebih baik dalam menangkap semua emosi dalam data, sehingga mengurangi risiko emosi hilang atau salah disajikan.

Analisis mendalam menunjukkan bahwa Naïve Bayes pandai menangani kumpulan data yang besar dan kompleks seperti data Twitter dan 13 Sentimen berbeda. Keuntungan utama Naïve Bayes adalah kemampuannya untuk menggabungkan hasil dari banyak teorema keputusan, memberikan prediksi yang stabil dan akurat. Hal ini sangat berguna dalam klasifikasi sentimen, dimana akurasi prediksi sangat penting untuk memahami sentimen publik.

Hasil penelitian ini menunjukkan bahwa Naïve Bayes lebih unggul dalam klasifikasi sentimen pada dataset yang besar dan kompleks seperti Twitter, dengan akurasi dan kinerja keseluruhan yang lebih baik dibandingkan Random Forest. Namun, pemilihan algoritma harus disesuaikan dengan karakteristik data dan kebutuhan analisis, di mana Random Forest dapat menjadi pilihan yang tepat untuk dataset yang lebih sederhana. Kedua algoritma ini memiliki tempatnya masing-masing dalam analisis sentimen media sosial, dan penggunaannya harus dipertimbangkan secara cermat berdasarkan tujuan spesifik penelitian.

SIMPULAN

Berdasarkan hasil penelitian, dapat disimpulkan bahwa algoritma Naïve Bayes menunjukkan keunggulan yang signifikan dalam klasifikasi sentimen, terutama pada dataset besar dan kompleks seperti yang ditemukan di Twitter, yang mencakup 13 jenis sentimen. Dengan tingkat akurasi mencapai 90,41%, Naïve Bayes berhasil memberikan prediksi yang akurat dan stabil, bahkan ketika dihadapkan pada data yang memiliki dimensi tinggi dan tidak terstruktur. Salah satu keunggulan utama dari Naïve Bayes adalah kemampuannya dalam mengurangi kesalahan prediksi yang merupakan aspek krusial dalam analisis

sentimen yang bertujuan untuk memahami secara tepat emosi atau pendapat. Selain itu, algoritma ini mampu menangani tantangan dari data besar dan kompleks dengan efisiensi tinggi, karena dapat menggabungkan berbagai teori keputusan untuk menghasilkan prediksi yang stabil dan responsif terhadap perubahan data.

Sebaliknya, algoritma Random Forest menunjukkan performa yang kurang memuaskan dengan akurasi sebesar 39,74%. Hal ini mengindikasikan bahwa algoritma ini kurang efektif saat dihadapkan pada dataset berdimensi tinggi dan kompleks. Meskipun Random Forest mampu beroperasi dengan baik pada dataset yang lebih sederhana, ia tidak mencapai tingkat efektivitas yang sama seperti Naïve Bayes dalam konteks analisis sentimen di media sosial. Namun, penting untuk diingat bahwa pemilihan algoritma harus disesuaikan dengan karakteristik data dan tujuan analisis yang ingin dicapai. Dalam beberapa kasus, terutama yang melibatkan dataset yang lebih sederhana, Random Forest masih bisa menjadi pilihan yang baik. Secara keseluruhan, Naïve Bayes terbukti lebih sesuai untuk aplikasi klasifikasi sentimen pada data besar dan kompleks, seperti yang ditemukan di Twitter, dengan memberikan akurasi dan kinerja yang lebih baik dibandingkan Random Forest.

DAFTAR PUSTAKA

A. Miftahusalam, A. F. Nuraini, A. A. Khoirunisa, and H. Pratiwi, "Perbandingan Algoritma Random Forest, Naïve Bayes, dan Support Vector Machine Pada Analisis Sentimen Twitter Mengenai Opini Masyarakat Terhadap Penghapusan Tenaga Honorer," *Semin. Nas. Off. Stat.*, vol. 2022, no. 1, pp. 563–572, 2022, doi: 10.34123/semnasoffstat.v2022i1.1410.

A. Musman, *Seni Berdamai Dengan Emosi*. Unicorn Publishing, 2019.

A. Wandani, "Sentimen Analisis Pengguna Twitter pada Event Flash Sale Menggunakan Algoritma K-NN, Random Forest, dan Naive Bayes," *J. Sains Komput. Inform. (J-SAKTI*, vol. 5, no. 2, pp. 651–665, 2021.

D. P. M. Artanti, "Syukur A Prihandono A and Setiadi DRIM, 2018 Analisa Sentimen Untuk Penilaian Pelayanan Situs Belanja Online Menggunakan Algoritma Naıve Bayes ...," *Nas. Sist. Inf*, pp. 8–9, 2018.

E. Rini Yulia and K. Solecha, "Implementasi Particle Swarm Optimization (PSO) pada Analysis Sentiment Review Aplikasi Trafi menggunakan Algoritma Naive Bayes (NB)," *J. Tek. Komput. AMIK BSI*, vol. 7, no. 1, pp. 25–29, 2021, doi: 10.31294/jtk.v4i2.

F. A. Larasati, D. E. Ratnawati, and B. T. Hanggara, "Analisis Sentimen Ulasan Aplikasi Dana dengan Metode Random Forest," *... Teknol. Inf. dan ...*, vol. 6, no. 9, pp. 4305–4313,

2022, [Online]. Available: http://j-ptiik.ub.ac.id

G. A. Sandag, "Prediksi Rating Aplikasi App Store Menggunakan Algoritma Random Forest," *CogITo Smart J.*, vol. 6, no. 2, pp. 167–178, 2020, doi: 10.31154/cogito.v6i2.270.167-178.

J. M. Ayu, S. Dachi, and P. Sitompul, "Analisis Perbandingan Algoritma XGBoost dan Algoritma Random Forest Ensemble Learning pada Klasifikasi Keputusan Kredit," *J. Ris. Rumpun Mat. dan Ilmu Pengetah. Alam*, vol. 2, no. 2, pp. 87–103, 2023, [Online]. Available: https://prin.or.id/index.php/JURRIMIPA/article/view/1470

M. H. Chyntia, D. E. Ratnawati, and I. Arwani, "Analisis Sentimen berbasis Aspek terhadap Ulasan Hotel Tentrem Yogyakarta menggunakan Algoritma Random Forest Classifier," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 6, no. 4, pp. 1702–1708, 2022, [Online]. Available: http://j-ptiik.ub.ac.id

M. Yasir and R. Suraji, "Perbandingan Metode Klasifikasi Naive Bayes, Decision, Tree, Random Forest Terhadap Analisis Sentimen Kenaikan Biaya Haji 2023 pada Media Sosial Youtube," *J. Cahaya Mandalika*, vol. 3, no. 2, pp. 180–192, 2023.

N. D. Pratidina and J. Mitha, "Dampak Penggunaan Media Sosial terhadap Interaksi Sosial Masyarakat: Studi Literature," *J. Ilm. Univ. Batanghari Jambi*, vol. 23, no. 1, p. 810, 2023, doi: 10.33087/jiubj.v23i1.3083.

N. Widjiyati, "Implementasi Algoritme Random Forest Pada Klasifikasi Dataset Credit Approval," *J. Janitra Inform. dan Sist. Inf.*, vol. 1, no. 1, pp. 1–7, 2021, doi: 10.25008/janitra.v1i1.118.

R. Fatmasari, V. M. Ayu, H. Anto, W. Gata, and L. D. Yulianto, "Analisis Sentimen Dalam Pengkategorian Komentar Youtube Terhadap Layanan Akademik dan Non-Akademik Universitas Terbuka Untuk Prediksi Kepuasan," *Build. Informatics, Technol. Sci.*, vol. 4, no. 2, pp. 395–404, 2022, doi: 10.47065/bits.v4i2.1738.

R. Istiyaningsih, "Pengaruh Postingan Sindiran Di Media Sosial Facebook Terhadap Sikap Emosional Para Ibu Di Desa Agung Batin Kecamatan Simpang Pematang Kabupaten Mesuji," Universitas Lampung Bandar Lampung, 2022. [Online]. Available: http://digilib.unila.ac.id/id/eprint/66920

S. P. Dewi, N. Nurwati, and E. Rahayu, "Penerapan Data Mining Untuk Prediksi Penjualan Produk Terlaris Menggunakan Metode K-Nearest Neighbor," *Build. Informatics, Technol. Sci.*, vol. 3, no. 4, pp. 639–648, 2022, doi: 10.47065/bits.v3i4.1408.

S. S. HANDAYANI, "Regulasi Emosi Pada Pengguna Media Sosial," Universitas Muhammadiyah Surakarta, 2018. [Online]. Available: https://eprints.ums.ac.id/id/eprint/61449

T. D. Putra, E. Utami, and M. P. Kurniawan, "Analisis Sentimen Pemilu 2024 dengan Naive Bayes Berbasis Particle Swarm Optimization (PSO)," *Explore*, vol. 13, no. 1, pp. 1–5, 2023, doi: 10.35200/ex.v11i2.13.

T. D. Putra, E. Utami, and M. P.Kurniawan, "Klasifikasi penderita kanker Paru Paru Menggunakan Algoritma Artificial Neural Network (ANN)," *Explore*, vol. 12, no. 2, p. 13, 2022, doi: 10.35200/explore.v12i2.568.

T. Fadiyah Basar, D. E. Ratnawati, and I. Arwani, "Analisis Sentimen Pengguna Twitter terhadap Pembayaran Cashless menggunakan Shopeepay dengan Algoritma Random Forest," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 6, no. 3, pp. 1426–1433, 2022, [Online]. Available: http://j-ptiik.ub.ac.id

T. Krisdiyanto, "Analisis Sentimen Opini Masyarakat Indonesia Terhadap Kebijakan PPKM pada Media Sosial Twitter Menggunakan Naïve Bayes Clasifiers," *J. CoreIT J. Has. Penelit. Ilmu Komput. dan Teknol. Inf.*, vol. 7, no. 1, p. 32, 2021, doi: 10.24014/coreit.v7i1.12945.