Institut Riset dan Publikasi Indonesia (IRPI)
MALCOM: Indonesian Journal of Machine Learning and Computer Science
Journal Homepage: https://journal.irpi.or.id/index.php/malcom
Vol. 5 Iss. 2 April 2025, pp: 566-577
ISSN(P): 2797-2313 | ISSN(E): 2775-8575

Application of the Naive Bayes Algorithm with TF-IDF and Cross Validation Techniques for Sentiment Analysis Towards Starlink

Penerapan Algoritma Naive Bayes dengan Teknik TF-IDF dan Cross Validation untuk Analisis Sentimen Terhadap Starlink

Suci Khoerunnisa[1*], Diqy Fakhrun Shiddiq[2], Dwi Nurhayati[3]

[1,2,3]Program Studi Bisnis Digital, Fakultas Ekonomi, Universitas Garut, Indonesia

E-Mail: [1]24025121066@fekon.uniga.ac.id, [2]diqy@uniga.ac.id, [3]dwinurhayati@uniga.ac.id

Received Dec 24th 2024; Revised Feb 14th 2025; Accepted Feb 22th 2025; Available Online Mar 21th 2025, Published Jan 21th 2025
Corresponding Author: Suci Khoerunnisa

Abstract

Starlink, a satellite internet service from SpaceX, will begin operations in Indonesia in 2024 to address the digital divide in remote areas. However, its presence poses challenges such as high prices, potential impact on local providers, and regulatory issues. This research analyzes public sentiment towards Starlink using Naïve Bayes algorithm combined with TF-IDF and Cross Validation techniques which are still rarely applied in similar studies in Indonesia. The data used are Indonesian tweets from users of platform X during May-November 2024. The analysis results show that the Naïve Bayes model has optimal performance in detecting positive sentiment compared to negative or neutral, as measured using confusion matrix. The main findings showed that Naïve Bayes 49.38% of tweets had positive sentiment, 32.94% were neutral, and 17.68% were negative. The positive sentiment was dominated by appreciation of the speed and stability of the service, while the negative sentiment criticized the high price and its impact on local providers. While the model performed well on positive sentiments, the classification accuracy of negative and neutral sentiments still needs to be improved. The results of this study provide strategic insights for Starlink's business development as well as a basis for consideration for the government regarding satellite-based internet services in Indonesia.

Keyword: Cross Validation, Naïve Bayes, Sentiment Analysis, Starlink, TF-IDF

Abstrak

Starlink, layanan internet satelit dari SpaceX, mulai beroperasidi Indonesia pada 2024 untuk mengatasi kesenjangan digital di wilayah terpencil. Namun, kehadirannya menimbulkan tantangan seperti harga tinggi, potensi dampak terhadap penyedia lokal, dan masalah regulasi. Penelitian ini mengkaji sentimen publik terhadap Starlink menggunakan algoritma Naïve Bayes yang dikombinasikan dengan teknik TF-IDF dan *Cross Validation* yang masih jarang diterapkan dalam studi serupa di Indonesia. Data yang digunakan berupa cuitan berbahasa Indonesia dari pengguna platform X selama Mei-November 2024. Hasil analisis menunjukkan bahwa model Naïve Bayes memiliki kinerja optimal dalam mendeteksi sentimen positif dibandingkan negatif maupun netral, sebagaimana diukur menggunakan confusion matrix. Temuan utama menunjukkan bahwa Naïve Bayes 49,38% cuitan bersentimen positif, 32,94% netral, dan 17,68% negatif. Sentimen positif didominasi oleh apresiasi terhadap kecepatan dan stabilitas layanan, sedangkan sentimen negatif mengkritik harga tinggi dan dampaknya terhadap penyedia lokal. Meskipun model menunjukkan performa baik pada sentimen positif, akurasi klasifikasi sentimen negatif dan netral masih perlu ditingkatkan. Hasil penelitian ini memberikan wawasan strategis bagi pengembangan bisnis Starlink serta dasar pertimbangan bagi pemerintah terkait layanan internet berbasis satelit di Indonesia.

Kata Kunci: Analisis Sentimen, Cross Validation, Naïve Bayes, Starlink, TF-IDF

1. PENDAHULUAN

Internet pertama kali dikembangkan oleh Advanced Research Projects Agency (ARPA) pada 1969 sebagai Advanced Research Projects Agency Network (ARPANET), kini menjadi elemen krusial dalam rutinitas harian [1][2]. Di Indonesia, jumlah pengguna internet pada 2024 diperkirakan mencapai 221,5 juta dengan penetrasi 79,5%, dipengaruhi oleh penyebaran infrastruktur jaringan yang merata, baik di perkotaan maupun wilayah rural [3][4]. Pada 19 Mei 2024, Starlink, layanan internet satelit dari SpaceX, resmi

diluncurkan di Indonesia dengan tujuan mengurangi kesenjangan digital, terutama di daerah terpencil [5]. Meskipun demikian, peluncuran Starlink menghadapi berbagai tantangan, termasuk regulasi ketat dan potensi dampak terhadap industri lokal [4]. Kehadiran Starlink juga memicu perbincangan publik, terutama di platform X (sebelumnya Twitter), yang menjadikan analisis sentimen di platform ini penting untuk memahami persepsi masyarakat terhadap layanan ini [6][7][8].

Analisis sentimen digunakan untuk mengevaluasi pandangan, sentimen, penilaian, sikap, serta emosi masyarakat terhadap entitas dan atributnya melalui teks mengenai berbagai topik, produk, subjek, dan layanan [9][10]. Naive Bayes (NB) merupakan metode klasifikasi yang banyak digunakan karena kesederhanaannya dan kinerja yang cukup baik [11]. Penelitian sebelumnya menunjukkan bahwa 68,99% dari 416 tweet tentang Starlink memiliki sentimen positif, dengan akurasi model sebesar 80% [12]. Penelitian lain menggunakan SVM untuk menganalisis 1.976 tweet tentang Starlink di Indonesia dan menemukan 56,3% tweet dengan sentimen positif, dengan akurasi 76,22% [13]. Sebuah studi yang membandingkan Naive Bayes dan SVM dalam analisis sentimen *cryptocurrency* menemukan bahwa Naive Bayes mencapai akurasi sebesar 77,94%, sementara SVM hanya 72,85% [14]. Penelitian lain yang menganalisis sentimen terhadap layanan Indihome dengan Naive Bayes mencapai akurasi 82% [15]. Adapun penelitian sebelumnya menggunakan analisis sentimen dengan metode Naïve Bayes untuk mengukur kepuasan pengguna terhadap layanan Gojek dan Grab di Indonesia, dengan hasil akurasi 99,80% untuk Gojek dan 99,90% untuk Grab, menunjukkan dominasi opini negatif dari pengguna di Twitter [16]. Namun, penelitian sebelumnya cenderung belum mengeksplorasi secara mendalam kombinasi teknik seperti penggunaan TF-IDF dan *cross-validation* bersama algoritma Naive Bayes dalam analisis sentimen layanan internet satelit. Penggunaan TF-IDF dapat menghasilkan klasifikasi yang lebih tepat, karena teknik ini memberikan perhatian lebih pada kata-kata yang dianggap penting dalam konteks data, sedangkan *cross-validation* digunakan untuk mengevaluasi model secara menyeluruh [17][18]. Gap ini penting untuk diisi karena kombinasi metode tersebut berpotensi menghasilkan klasifikasi sentimen publik dengan tingkat akurasi yang lebih tinggi.

Urgensi dari penelitian ini dilatarbelakangi oleh kebutuhan mendesak untuk memahami persepsi publik terhadap layanan internet satelit, seperti Starlink, yang semakin populer namun juga kontroversial di berbagai kalangan. Sebagai layanan yang relatif baru dan disruptif, Starlink menghadapi berbagai opini yang beragam di media sosial, yang dapat mempengaruhi persepsi pengguna potensial dan kebijakan publik terkait teknologi satelit [19]. Tujuan penelitian ini adalah menerapkan algoritma Naive Bayes dengan teknik TF-IDF dan Cross Validation dalam analisis sentimen terkait kehadiran Starlink di Indonesia. Penelitian ini diharapkan dapat mengungkapkan opini publik mengenai Starlink serta mengevaluasi efektivitas metode yang digunakan. Hasilnya akan memberikan wawasan penting untuk pengembangan dan penerapan analisis sentimen yang lebih baik dalam mengklasifikasi opini publik terhadap kehadiran teknologi baru seperti Starlink. Bagi Starlink, penelitian ini dapat mendukung pengambilan keputusan bisnis yang lebih baik di sektor bisnis dan mitigasi risiko reputasi.

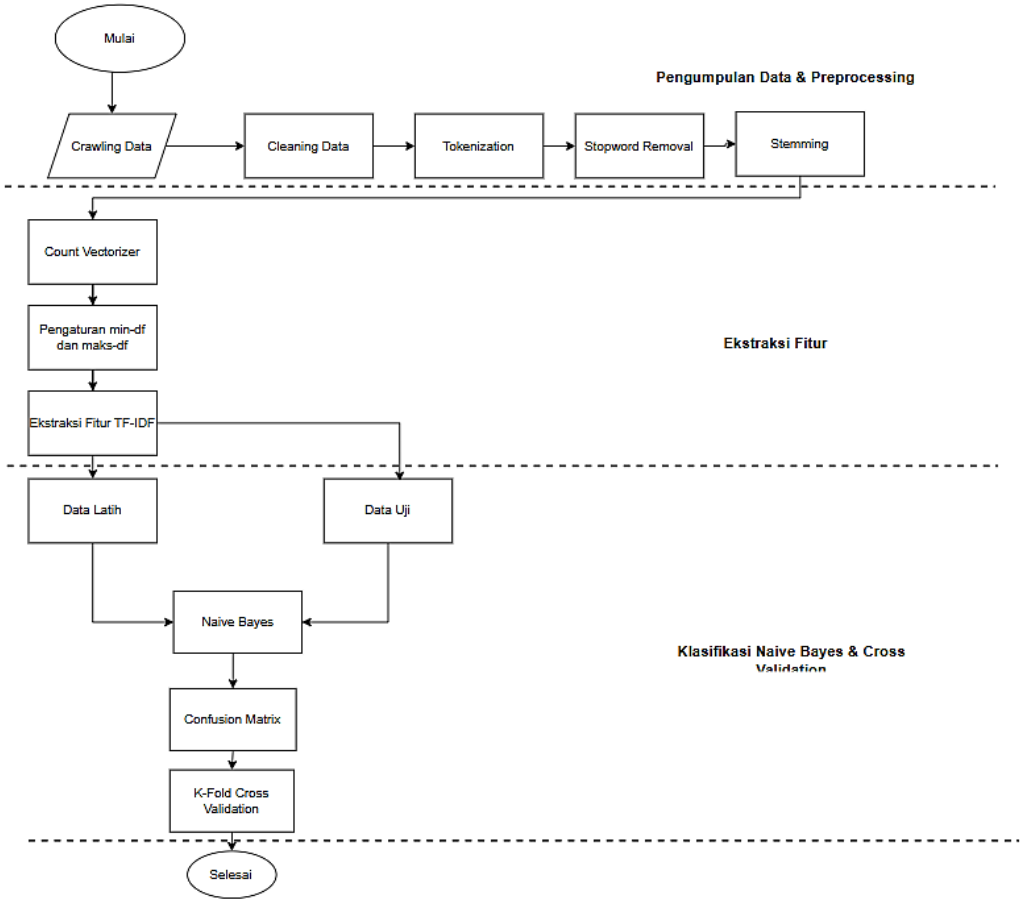
2. STUDI LITERATUR

Penelitian mengenai analisis sentimen menggunakan teknik pembelajaran mesin, seperti *Support Vector Machine* (SVM), Naive Bayes, dan *Deep Learning*, telah berkembang pesat, terutama dalam mempelajari pandangan publik melalui media sosial mengenai produk dan layanan teknologi. Salah satu pendekatan yang populer adalah Naive Bayes, yang dikenal dengan kesederhanaannya dan efektivitasnya dalam klasifikasi teks [9][11]. Beberapa penelitian, seperti yang dilakukan oleh Gibran et al. (2024) mengenai sentimen pengguna terhadap Starlink di Twitter, menunjukkan hasil yang signifikan dengan 68,99% tweet berisi sentimen positif dan akurasi model sebesar 80% [11]. Penelitian lainnya oleh Nugroho dan Kurniadi (2024) menggunakan SVM, yang menemukan 56,3% tweet mengenai Starlink dengan sentimen positif, dan akurasi model sebesar 76,22%[11]. Adapun penelitian lain sebelumnya menganalisis sentimen pengguna terhadap layanan Gojek dan Grab di Indonesia menggunakan metode Naïve Bayes, dengan hasil akurasi 99,80% untuk Gojek dan 99,90% untuk Grab. Hasil analisis menunjukkan bahwa mayoritas opini di Twitter mengenai kedua layanan ini bersifat negatif [16].

Teknik TF-IDF juga terbukti meningkatkan kinerja model klasifikasi sentimen, memberikan prioritas pada kata-kata penting dalam dokumen untuk membantu model memahami nuansa sentimen dalam teks [17][18]. Dalam analisis sentimen mengenai *cryptocurrency*, Naive Bayes menunjukkan akurasi yang lebih tinggi (77,94%) dibandingkan dengan SVM (72,85%) [14]. Penelitian oleh Permataning Tyas et al. (2022) membuktikan bahwa algoritma Naïve Bayes memperoleh nilai akurasi sebesar 82% dalam proses analisis sentimen terhadap layanan Indihome, menegaskan efektivitas algoritma ini dalam konteks teknologi komunikasi [14]. Namun, sebagian besar penelitian belum menggabungkan teknik TF-IDF dan *Cross-validation* secara mendalam bersama Naive Bayes, terutama dalam menganalisis layanan internet satelit seperti Starlink. Penelitian lebih lanjut diharapkan dapat meningkatkan akurasi dalam mengklasifikasikan sentimen publik terhadap teknologi baru seperti Starlink, yang menghadapi tantangan regulasi dan persaingan di Indonesia[7][14][20].

3. METODOLOGI PENELITIAN

Penelitian ini memanfaatkan algoritma Naïve Bayes dengan teknik TF-IDF dan *Cross Validation* untuk melakukan analisis sentimen terhadap Layanan Internet Starlink. Pada gambar 1 terdapat alur penelitian.



Gambar 1. Alur Penelitian

2.1. *Crawling* Data

Crawling data yang diambil dari X melibatkan pengambilan data pengguna dan tweet yang diambil berdasarkan kata kunci spesifik. Proses ekstraksi ini dilakukan dengan memanfaatkan *Application Programming Integration* (API) [21]. Dalam penelitian ini, *crawling* dilakukan menggunakan pustaka tweet-harvest dari bahasa pemrograman Python. Pengumpulan data difokuskan hanya pada tweet berbahasa Indonesia selama periode tiga bulan, dari 19 Mei 2024 hingga 04 November 2024, dengan jumlah tweet 10003 yang menggunakan kata kunci Starlink.

2.2. *Preprocessing* Data

Langkah-langkah *preprocessing* teks bertujuan untuk mengolah dan membersihkan teks sehingga siap dipakai dalam proses klasifikasi [22]. *Preprocessing* data penting dilakukan untuk memahami dan mengenali permasalahan yang terkait dengan data, sehingga data menjadi lebih andal dan siap digunakan [23]. Langkah-langkah yang diterapkan dalam tahap ini meliputi: [24].

1. *Case folding* adalah teknik yang mengubah seluruh teks seluruhnya menjadi huruf kecil.
2. *Convert* Emoji dimana emoji diubah menjadi kata-kata yang merepresentasikan arti dari emoji tersebut.
3. *Cleansing* merujuk pada proses pembersihan dokumen dari elemen-elemen seperti URL, angka, karakter simbol, serta tanda baca lainnya.
4. *Tokenizing* merupakan proses pemecahan atau pemisahan teks, baik itu dokumen atau kalimat, menjadi unit - unit kata.
5. Normalisasi kata merupakan langkah mengubah kata-kata yang tidak mengikuti aturan bahasa resmi atau mengandung gangguan menjadi bentuk yang standar dan dapat diolah. Gangguan ini bisa berupa kata-kata yang menggunakan bahasa daerah yang tidak selaras dengan KBBI atau singkatan yang lazim di media sosial.. f. Stopword removal merupakan metode untuk menghilangkan kata-kata yang

dianggap tidak penting, yakni kata - kata yang tidak membawa makna signifikan, berdasarkan daftar kata yang telah dikategorikan sebagai non-informatif.

6. Stemming adalah proses mengubah kata-kata dalam teks menjadi bentuk dasar dengan cara menghapus imbuhan.

2.3. Pembobotan Kata TF-IDF

TF-IDF memberikan bobot pada kata-kata berdasarkan frekuensinya dalam sebuah dokumen relatif terhadap frekuensinya di seluruh dokumen dalam korpus . Kata yang signifikan diberi bobot lebih tinggi. Sebaliknya, jika kata umum digunakan di seluruh kategori, bobotnya perlu diturunkan meskipun frekuensinya tinggi di beberapa kumpulan dokumen [25]. Hitung banyaknya dokumen yang memuat kata (DF) kemudian hitung IDF dengan persamaan 1 [26] dan 2 [27].

$$IDF(w) = log((\frac{N}{\mathrm{DF(W)}}) \quad (1)$$

$$w(\mathrm{t,d}) = t(f, d) \times idf(t)) \quad (2)$$

wt,d mewakili bobot istilah (t) dalam dokumen (d), sementara tff,d merujuk padafrekuensi kemunculan istilah (t) dalam dokumen (d), dan idft adalah nilai frekuensi terbalik dokumen.

2.4. Naïve Bayes

Naïve Bayes merupakan salah satu algoritma yang digunakan dalam klasifikasi multikelas. Nama "Naïve Bayes" mengacu pada penyederhanaan perhitungan probabilitas untuk setiap kategori, sehingga proses komputasi menjadi lebih efisien. Algoritma ini dikembangkan berdasarkan pendekatan klasifikasi Bayesian, yang bergantung pada penerapan teorema Bayes. Teorema ini menjelaskan hubungan antara probabilitas bersyarat dalam suatu kumpulan data, memungkinkan kita untuk menafsirkannya dalam bentuk nilai yang lebih mudah dihitung secara langsung [28]. Persamaan umum Naïve Bayes ditunjukkan pada persamaan 3 [29].

$$P(U/V) = \frac{p(V/U)P(U)}{P(V)} \quad (3)$$

2.5. *Cross Validation*

K-Fold Cross Validation adalah teknik pengujian yang membagi data ke dalam beberapa subset untuk mengurangi bias dalam proses pelatihan dan pengujian. Teknik ini sering digunakan karena terbukti efektif dalam memastikan bahwa model tidak hanya terlatih dengan baik pada satu kelompok data, tetapi juga pada keseluruhan dataset [30]. Untuk mengukur tingkat akurasi model klasifikasi yang digunakan, diperlukan metode validasi pengujian, yaitu *K-Fold Cross Validation*, karena jumlah dataset yang cukup besar. Untuk membagi data menjadi bagian pelatihan dan pengujian dari dataset tersebut, digunakan metode *K-Fold Cross Validation*. *K-Fold Cross Validation* adalah metode uji yang membagi seluruh data menjadi data latih dan data uji [31]. Untuk melihat kinerja setiap Naive Bayes, penulis menetapkan nilai K=10 pada *K-Fold Cross Validation* dengan 10 iterasi pelatihan dan pengujian. Hasil pengujian dari setiap iterasi pada *K-Fold Cross Validation* untuk algoritma Naive Bayes dievaluasi secara menyeluruh. Evaluasi model dalam penelitian ini memanfaatkan metode confusion matrix untuk menghitung akurasi serta menggambarkan performa model klasifikasi [32].

3. HASIL DAN PEMBAHASAN

3.1. *Crawling* Data

Proses *crawling* data merupakan langkah awal yang perlu diterapkan dalam analisis sentimen. Pada tahap pengumpulan data, informasi diperoleh melalui platform X. Data yang dikumpulkan meliputi ulasan dari pengguna, dengan rentang waktu pengambilan data antara 19 Mei 2024 hingga 04 November 2024, dan jumlah data yang berhasil terkumpul sebanyak 10.003.

Tabel 1. Tampilan Hasil Crawling Data

No	Cuitan
1	@IndiHomeCare Makin gila nih Indihome download bukan jam lagi hari... waktunya nyoba provider lain lah... buat yg d pelosok saranin coba yg baru aje starlink temen pake d pelosok bagus .... untung dah akhir bulan bye bye indihome...
2	@anggaandinata menurut penasehat spiritualnya krn dipandang bahwa rakyat sangat menikmati kercep tol starlink dokter aseng IKN dsb dsb..maka rakyat pasti mau bayar segala kompensasinya @PNS_Ababil @PartaiSocmed

No	Cuitan
…	
1002	Kang @farhanpenyiar bagi2 wifi gratis aja pk starlink +wifi extender watt gede radius 1km. Landing page berpassword jargon reroute dl ke web profil kayak di hotel2. Dgn 30menit hrs rekonek 200+mbps bs 300 koneksi/jam. Janji prmanen kl menang. Smp coblosan plg modal 10jt/titik
1003	Pernah ga kalian stream trus indihumu jelek banget saking jeleknya ada someone yg cek perkiraan wifi starlink di indonesia trus kirim uang buat beli wifi https://t.co/NFUjuk66Gf

Tabel 1 menunjukkan contoh tampilan hasil dari proses pengumpulan data, yang mencakup teks lengkap dari cuitan.

3.2. *Preprocessing* Data

Setelah data terkumpul, langkah berikutnya adalah melakukan pra-pemrosesan agar data siap digunakan. Berikut merupakan beberapa metode pra pemrosesan data yang diterapkan.

3.2.1 *Cleaning* Data

Data yang berhasil diperoleh melalui proses crawling selanjutnya akan diproses dalam tahap pembersihan untuk menghapus elemen-elemen seperti URL, emotikon, *retweet*, simbol, hashtag, serta spasi yang tidak diperlukan. Selain itu, cuitan yang menggunakan bahasa selain bahasa Indonesia dan yang tidak relevan juga akan dibersihkan. Tabel 2 menunjukkan perbandingan antara data sebelum dan setelah melalui tahap pembersihan.

Tabel 2. Cleaning Data

No	Sebelum	Sesudah
1	@IndiHomeCare Makin gila nih Indihome download bukan jam lagi hari... waktunya nyoba provider lain lah... buat yg d pelosok saranin coba yg baru aje starlink temen pake d pelosok bagus .... untung dah akhir bulan bye bye indihome...	makin gila nih indihome download bukan jam lagi hari waktunya nyoba provider lain lah buat yg d pelosok saranin coba yg baru aje starlink temen pake d pelosok bagus untung dah akhir bulan bye bye indihome
2	@Leonita_Lestari Tapi saya dukung starlink beroperasi di indonesia. Sdh lama kita dijadikan mainan para provider dgn kualitas amburadul. Kalau provider tdk berubah ya sebaiknya hancur saja. Langganan 750rb/bln tnpa fup dan speed >150Mbps apa provider k	tapi saya dukung starlink beroperasi di indonesia sdh lama kita dijadikan mainan para provider dgn kualitas amburadul kalau provider tdk berubah ya sebaiknya hancur saja langganan rbbln tnpa fup dan speed gtmbps apa provider k
…		
6291	Pernah ga kalian stream trus indihumu jelek banget saking jeleknya ada someone yg cek perkiraan wifi starlink di indonesia trus kirim uang buat beli wifi https://t.co/NFUjuk66Gf	pernah ga kalian stream trus indihumu jelek banget saking jeleknya ada someone yg cek perkiraan wifi starlink di indonesia trus kirim uang buat beli wifi

3.2.2 *Tokenization*

Tahap tokenisasi adalah proses memecah keseluruhan teks ulasan menjadi unit kata yang lebih kecil (disebut token). Tokenisasi dapat berupa unit linguistik kata-kata, frasa, kalimat, atau karakter, dengan tujuan mempermudah tahapan pemrosesan teks untuk analisis lebih lanjut. Tabel 3 menunjukkan perbandingan antara data sebelum dan setelah melalui tahap tokenisasi, di mana kata-kata telah dipisah-pisah.

Tabel 3. Tokenization

No	Sebelum	Sesudah
1	makin gila nih indihome download bukan jam lagi hari waktunya nyoba provider lain lah buat yg d pelosok saranin coba yg baru aje starlink temen pake d pelosok bagus untung dah akhir bulan bye bye indihome	['makin', 'gila', 'nih', 'indihome', 'download', 'bukan', 'jam', 'lagi', 'hari', 'waktunya', 'nyoba', 'provider', 'lain', 'lah', 'buat', 'yg', 'd', 'pelosok', 'saranin', 'coba', 'yg', 'baru', 'aje', 'starlink', 'temen', 'pake', 'd', 'pelosok', 'bagus', 'untung', 'dah', 'akhir', 'bulan', 'bye', 'bye', 'indihome']
2	tapi saya dukung starlink beroperasi di indonesia sdh lama kita dijadikan mainan para provider dgn kualitas amburadul kalau provider tdk berubah ya sebaiknya hancur saja langganan rbbln tnpa fup dan speed gtmbps apa provider k	['tapi', 'saya', 'dukung', 'starlink', 'beroperasi', 'di', 'indonesia', 'sdh', 'lama', 'kita', 'dijadikan', 'mainan', 'para', 'provider', 'dgn', 'kualitas', 'amburadul', 'kalau', 'provider', 'tdk', 'berubah', 'ya', 'sebaiknya', 'hancur', 'saja', 'langganan', 'rbbln', 'tnpa', 'fup', 'dan', 'speed', 'gtmbps', 'apa', 'provider', 'k']
…		
6291	pernah ga kalian stream trus indihumu jelek banget saking jeleknya ada someone yg cek perkiraan wifi	['pernah', 'ga', 'kalian', 'stream', 'trus', 'indihumu', 'jelek', 'banget', 'saking', 'jeleknya', 'ada', 'someone', 'yg', 'cek',

No	Sebelum	Sesudah
	starlink di indonesia trus kirim uang buat beli wifi	'perkiraan', 'wifi', 'starlink', 'di', 'indonesia', 'trus', 'kirim', 'uang', 'buat', 'beli', 'wifi']

3.2.3 *Stopword Removal*

Proses ini merupakan tahap eliminasi kata-kata yang tidak memberikan kontribusi berarti atau kurang relevan yang dapat mempengaruhi analisis sentimen, seperti "dan", "atau", "di", dan lainnya. Tabel 4 menunjukkan perbandingan data sebelum dan setelah dilakukan tahap penghapusan stopword.

Tabel 4. Stopword Removal

No	Sebelum	Sesudah
1	['makin', 'gila', 'nih', 'indihome', 'download', 'bukan', 'jam', 'lagi', 'hari', 'waktunya', 'nyoba', 'provider', 'lain', 'lah', 'buat', 'yg', 'd', 'pelosok', 'saranin', 'coba', 'yg', 'baru', 'aje', 'starlink', 'temen', 'pake', 'd', 'pelosok', 'bagus', 'untung', 'dah', 'akhir', 'bulan', 'bye', 'bye', 'indihome']	['gila', 'nih', 'indihome', 'download', 'jam', 'nyoba', 'provider', 'yg', 'd', 'pelosok', 'saranin', 'coba', 'yg', 'aje', 'starlink', 'temen', 'pake', 'd', 'pelosok', 'bagus', 'untung', 'dah', 'bye', 'bye', 'indihome']
2	['tapi', 'saya', 'dukung', 'starlink', 'beroperasi', 'di', 'indonesia', 'sdh', 'lama', 'kita', 'dijadikan', 'mainan', 'para', 'provider', 'dgn', 'kualitas', 'amburadul', 'kalau', 'provider', 'tdk', 'berubah', 'ya', 'sebaiknya', 'hancur', 'saja', 'langganan', 'rbbln', 'tnpa', 'fup', 'dan', 'speed', 'gtmbps', 'apa', 'provider', 'k'	['dukung', 'starlink', 'beroperasi', 'indonesia', 'sdh', 'dijadikan', 'mainan', 'provider', 'dgn', 'kualitas', 'amburadul', 'provider', 'tdk', 'berubah', 'ya', 'hancur', 'langganan', 'rbbln', 'tnpa', 'fup', 'speed', 'gtmbps', 'provider', 'k']
…		
6291	['pernah', 'ga', 'kalian', 'stream', 'trus', 'indihumu', 'jelek', 'banget', 'saking', 'jeleknya', 'ada', 'someone', 'yg', 'cek', 'perkiraan', 'wifi', 'starlink', 'di', 'indonesia', 'trus', 'kirim', 'uang', 'buat', 'beli', 'wifi']	['ga', 'stream', 'trus', 'indihumu', 'jelek', 'banget', 'saking', 'jeleknya', 'someone', 'yg', 'cek', 'perkiraan', 'wifi', 'starlink', 'indonesia', 'trus', 'kirim', 'uang', 'beli', 'wifi']

3.2.4 *Stemming*

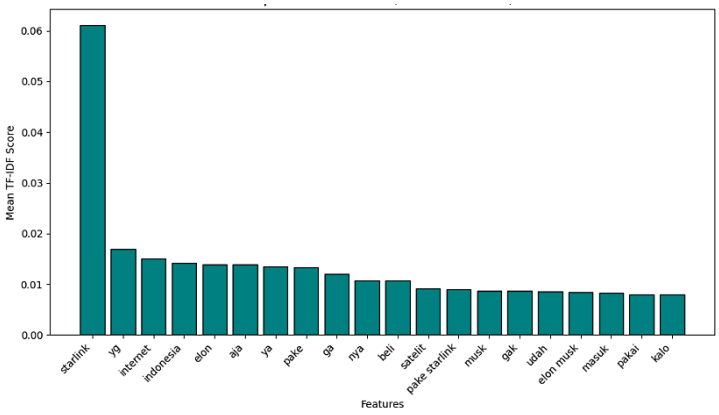
Stemming merupakan tahapan yang bertujuan untuk mengubah seluruh ulasan ke dalam bentuk kata dasar, dengan tujuan untuk menyederhanakan berbagai bentuk kata yang berasal dari kata dasar yang serupa. Sebagai contoh, kata-kata seperti "menghubungkan", "penghubung", dan "terhubung" akan diubah menjadi bentuk dasar "hubung". Berikut adalah Tabel 5 yang memperlihatkan perbandingan data sebelum dan sesudah dilakukan proses stemming.

Tabel 5. Stemming

No	Sebelum	Sesudah
1	['gila', 'nih', 'indihome', 'download', 'jam', 'nyoba', 'provider', 'yg', 'd', 'pelosok', 'saranin', 'coba', 'yg', 'aje', 'starlink', 'temen', 'pake', 'd', 'pelosok', 'bagus', 'untung', 'dah', 'bye', 'bye', 'indihome']	gila nih indihome download jam nyoba provider yg d pelosok saranin coba yg aje starlink temen pake d pelosok bagus untung dah bye bye indihome
2	['tapi', 'saya', 'dukung', 'starlink', 'beroperasi', 'di', 'indonesia', 'sdh', 'lama', 'kita', 'dijadikan', 'mainan', 'para', 'provider', 'dgn', 'kualitas', 'amburadul', 'kalau', 'provider', 'tdk', 'berubah', 'ya', 'sebaiknya', 'hancur', 'saja', 'langganan', 'rbbln', 'tnpa', 'fup', 'dan', 'speed', 'gtmbps', 'apa', 'provider', 'k'	dukung starlink operasi indonesia sdh jadi main provider dgn kualitas amburadul provider tdk ubah ya hancur langgan rbbln tnpa fup speed gtmbps provider k
…		
6291	['ga', 'stream', 'trus', 'indihumu', 'jelek', 'banget', 'saking', 'jeleknya', 'someone', 'yg', 'cek', 'perkiraan', 'wifi', 'starlink', 'indonesia', 'trus', 'kirim', 'uang', 'beli', 'wifi']	ga stream trus indihumu jelek banget saking jelek someone yg cek kira wifi starlink indonesia trus kirim uang beli wifi

3.3. Ekstraksi Fitur TF-IDF

Setelah data melalui tahap pra-pemrosesan, langkah selanjutnya adalah ekstraksi fitur menggunakan TF-IDF untuk mengonversi dataset ke dalam representasi vektor, yang dilakukan dengan bantuan teknik *CountVectorizer* dalam pustaka Python. TF-IDF, yang merupakan singkatan dari *Term Frequency* (TF) - *Inverse Document Frequency* (IDF) adalah teknik yang digunakan untuk menilai makna dari kalimat yang terdiri dari kata-kata.

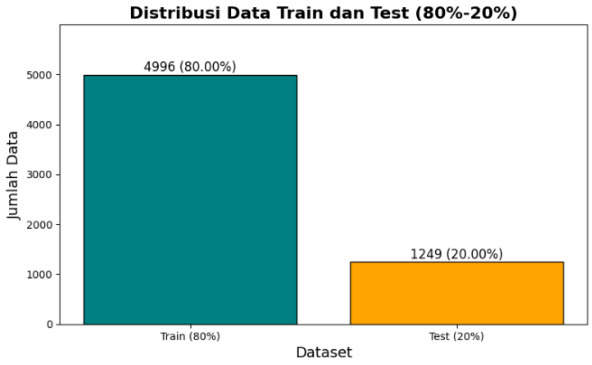


Gambar 2. Top 20 TF-IDF Features

Hasil visualisasi TF-IDF pada gambar 2 menunjukkan bahwa kata-kata dengan skor rata-rata tertinggi, seperti "Starlink", "yg", "internet", "indonesia", dan "elon", menjadi fitur utama yang berkontribusi signifikan dalam analisis sentimen terkait layanan Starlink. Kata "Starlink" memiliki skor tertinggi, mencerminkan fokus utama dokumen terhadap topik penelitian, sementara kata-kata lain seperti "pake" dan "beli" mencerminkan diskusi mengenai penggunaan atau pembelian layanan tersebut. Skor TF-IDF yang tinggi menandakan kata-kata ini lebih sering muncul dalam dokumen tertentu tetapi jarang di seluruh korpus, sehingga menjadi fitur penting dalam membedakan dokumen berdasarkan sentimen positif, negatif, maupun netral. Hal ini menunjukkan bahwa TF-IDF berhasil menangkap kata-kata kunci yang relevan dengan konteks penelitian dan membantu model klasifikasi, seperti Naive Bayes, dalam mengidentifikasi pola sentimen secara efektif.

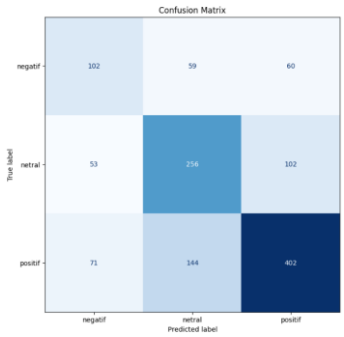
3.4. Klasifikasi Naïve Bayes

Sebelum model dikembangkan, dataset akan dipisahkan menjadi dua bagian yaitu data pelatihan dan data pengujian dengan rasio 80:20. Artinya, sebanyak 80% dari total dataset akan digunakan untuk pelatihan, sementara 20% sisanya dialokasikan sebagai data pengujian. Pada Gambar 3, jumlah data latih yang digunakan akan berjumlah 4996, sedangkan jumlah data uji akan berjumlah 1249.



Gambar 3. Distribusi Data

Performa model Naïve Bayes dianalisis melalui confusion matrix yang disajikan dalam gambar 4.



Gambar 4. Confusion Matrix

Gambar 4, *Confusion Matrix* ditampilkan beserta jumlah parameter yang dihasilkan seperti akurasi, presisi, *recall*, dan *f1-score*, yang diukur berdasarkan data. Hasil perhitungan dari parameter-parameter tersebut dapat dilihat pada gambar berikut.

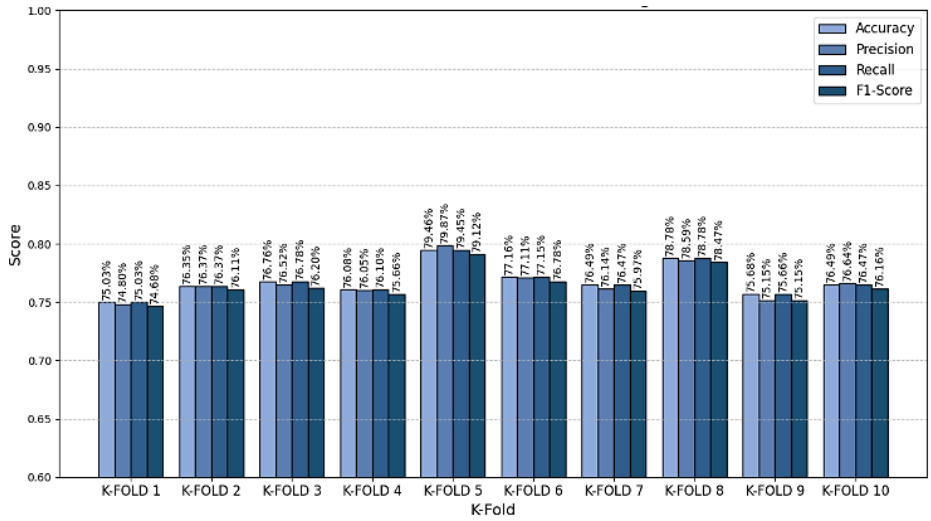
Tabel 6. Classification Report

	Precision	Recall	F1-score	Support
Negatif	0.45	0.46	0.46	221
Netral	0.56	0.62	0.59	411
Positif	0.71	0.65	0.68	617
Accuracy			0.61	1249
Macro Avg	0.57	0.58	0.58	1249
Weighted Avg	0.62	0.61	0.61	1249

Tabel 6 menampilkan hasil evaluasi performa model berdasarkan kolom Presisi. Nilai presisi untuk sentimen negatif tercatat sebesar 45%, untuk sentimen netral sebesar 56%, dan untuk sentimen positif sebesar 71%. Presisi ini menggambarkan sejauh mana model mampu secara akurat mengidentifikasi sentimen yang benar. Sementara itu, nilai *recal*l untuk sentimen negatif adalah 46%, sentimen netral 62%, dan sentimen positif mencapai 65%, yang menunjukkan sejauh mana model dapat mengenali teks dengan sentimen tertentu. Untuk *F1- Score*, sentimen negatif memperoleh nilai 46%, netral 59%, dan positif 68%. Hasil ini menunjukkan bahwa model lebih efektif dalam mengenali sentimen positif dibandingkan sentimen negatif maupun netral. Hal ini terlihat dari nilai presisi, *recall*, dan *F1-Score* yang lebih tinggi untuk sentimen positif. Rendahnya recall untuk sentimen negatif menunjukkan bahwa model memiliki keterbatasan dalam mengenali sentimen tersebut, sehingga secara keseluruhan pengguna Twitter cenderung menunjukkan respons yang lebih positif terhadap Starlink.

3.5. *Cross Validation*

Berdasarkan hasil pengujian yang dilakukan dengan menerapkan algoritma Naïve Bayes, dilakukan pengujian lagi dengan teknik *Cross Validation* menggunakan K-Folds agar mendapatkan hasil validasi yang lebih akurat pada gambar 5.



Gambar 5. 10 *Fold Cross Validation Testing*

Hasil visualisasi *10-Fold Cross Validation* pada gambar 5 menunjukkan bahwa model memiliki performa yang stabil di setiap fold. Rata-rata akurasi berkisar antara 76% hingga 79%, dengan fold terbaik adalah K-Fold 5 dan K-Fold 8, yang mencapai akurasi mendekati 79%. Akurasi yang konsisten di berbagai fold menunjukkan bahwa model memiliki kemampuan generalisasi yang cukup baik dan tidak *overfitting* pada subset data tertentu. Precision, recall, dan F1-score juga menunjukkan pola yang konsisten di setiap fold, dengan rata-rata nilai berada di sekitar 76% hingga 78%. *F1- score* sebagai rata-rata harmonik dari *precision* dan *recall*, memberikan gambaran keseimbangan antara keduanya, yang penting dalam evaluasi model klasifikasi. Hal ini mengindikasikan bahwa model mampu mempertahankan performanya secara konsisten dalam berbagai subset data, menunjukkan generalisasi yang baik. Stabilitas performa di berbagai fold menunjukkan bahwa model berhasil menghindari *overfitting*, yang merupakan indikator kemampuan generalisasi model pada data baru.

3.6. *Word Cloud*

Gambar 6. *Word Cloud* Sentimen Positif

Gambar 6 menampilkan hasil dari analisis sentimen positif terhadap topik "Starlink" dalam konteks Indonesia. Kata-kata yang sering muncul seperti "Starlink", "internet", "Indonesia", "harga", "pake", dan "murah" menunjukkan bahwa diskusi positif terkait Starlink sebagian besar berpusat pada aspek manfaat teknologi ini untuk akses internet di Indonesia. Kata "cepat", "stabil", dan "sinyal" menegaskan bahwa Starlink dilihat sebagai solusi untuk meningkatkan kualitas konektivitas, khususnya di daerah yang sulit dijangkau oleh jaringan tradisional.

Gambar 7. *Word Cloud* Sentimen Negatif

Gambar 7 menggambarkan hasil analisis sentimen negatif terhadap topik "Starlink" dalam konteks Indonesia. Kata-kata yang dominan, seperti "Starlink", "internet", "Indonesia", "harga", dan "mahal", menunjukkan bahwa sentimen negatif sebagian besar berpusat pada isu harga yang dianggap tidak terjangkau oleh sebagian masyarakat. Kata "blokir", "takut", dan "ancam" mencerminkan kekhawatiran terhadap dampak Starlink, baik dari segi persaingan dengan penyedia lokal maupun implikasi terhadap kedaulatan digital negara. Kemunculan kata "pemerintah", "kominfo", dan "negara" menunjukkan adanya kritik terhadap peran pemerintah dalam menyikapi kehadiran Starlink. Hal ini dapat mengindikasikan ketidakpuasan atau ketidakpercayaan masyarakat terhadap kesiapan pemerintah dalam mengatur atau mendukung implementasi teknologi ini secara adil.

Gambar 8. *Word Cloud* Sentimen Netral

Gambar 8 menunjukkan hasil dari analisis sentimen netral terhadap topik "Starlink" dalam diskusi di Indonesia. Kata-kata dominan seperti "Starlink", "internet", "Indonesia", "satelit", dan "harga" mencerminkan fokus utama percakapan yang bersifat informatif atau faktual tanpa adanya kecenderungan emosi yang signifikan, baik positif maupun negatif. Hal ini menunjukkan bahwa banyak diskusi terkait Starlink yang berorientasi pada eksplorasi informasi atau pengamatan terhadap layanan ini. Kata "pake", "mbps", dan "paket" menunjukkan perhatian masyarakat pada aspek teknis dan penawaran layanan Starlink. Diskusi ini

mengindikasikan bahwa pengguna cenderung mendiskusikan spesifikasi dan kepraktisan teknologi, seperti kecepatan internet atau harga paket yang ditawarkan, dalam konteks yang lebih netral.

3.7. Diskusi

Penelitian ini mengkaji sentimen publik terhadap layanan Starlink di Indonesia menggunakan algoritma Naïve Bayes yang dipadukan dengan TF-IDF dan *Cross Validation*. Hasil menunjukkan dominasi sentimen positif (49,38%), diikuti oleh sentimen netral (32,94%) dan negatif (17,68%). Temuan ini mengindikasikan bahwa meskipun ada kritik terkait harga yang tinggi dan dampaknya terhadap penyedia layanan lokal, secara keseluruhan, publik cenderung memiliki persepsi positif terhadap Starlink. Hal ini sejalan dengan temuan dari penelitian sebelumnya yang ditemukan oleh Gibran et al. (2024), menunjukkan kecenderungan positif terhadap Starlink yaitu sebesar 68,99% meskipun ada isu terkait harga dan dampaknya terhadap penyedia layanan lokal [12]. Penelitian ini juga menemukan bahwa meskipun model ini efektif dalam mengidentifikasi sentimen positif, akurasi dalam mengklasifikasikan sentimen negatif dan netral masih memerlukan peningkatan.

Teknik *Term Frequency-Inverse Document Frequency* (TF-IDF) dan *Cross Validation* memainkan peran penting dalam meningkatkan ketepatan model klasifikasi sentimen. TF-IDF membantu model lebih fokus pada kata-kata bermakna dengan memberikan bobot lebih tinggi pada kata yang signifikan, sementara kata-kata umum yang kurang informatif diberi bobot lebih rendah. Hal ini sejalan dengan penelitian Addiga dan Bagui (2022) serta Zhang et al. (2020) yang menunjukkan bahwa penerapan TF-IDF dalam analisis sentimen berhasil menangkap kata-kata kunci yang relevan dengan konteks penelitian [17][18]. Selain itu, metode *K-Fold Cross Validation* dengan K=10 diterapkan untuk memastikan performa model tetap stabil dan tidak mengalami *overfitting*. Hasil pengujian menunjukkan tingkat akurasi rata-rata 76%-79%, dengan performa terbaik pada fold ke-5 dan ke-8 yang mencapai hampir 79%. Temuan ini selaras dengan penelitian Ridwansyah (2022), yang menunjukkan bahwa *Cross Validation* meningkatkan keandalan model dengan mengurangi bias distribusi data. Hal ini juga didukung oleh penelitian Nugroho dan Kurniadi (2024), yang menemukan bahwa kombinasi teknik ini mampu memberikan hasil lebih akurat dalam analisis sentimen di media sosial [13].

Keunggulan dalam penelitian ini adalah penerapan kombinasi teknik TF-IDF dan Cross Validation, dimana TF-IDF memberikan bobot lebih tinggi pada kata – kata kunci yang relevan, seperti “Starlink”, “internet”, dan “Indonesia”, sehingga model dapat lebih efektif dalam menangkap pola sentimen yang terkandung dalam tweet. Selain itu, Cross Validation dengan teknik K-Fold membantu memastikan bahwa modle memiliki kemampuan generalisasi yang baik, dengan hasil akurasi rata – rata 76%-79%. Hal ini menunjukkan bahwa model dapat beradaptasi dengan berbagai subset data tanpa mengalami overfitting.

Keterbatasan penelitian ini terletak pada ketidakseimbangan data, dengan dominasi sentimen positif yang lebih tinggi. Ketidakseimbangan ini dapat mempengaruhi efektivitas model dalam mengklasifikasikan sentimen negatif dan netral, yang perlu diperbaiki di masa depan dengan model yang lebih kompleks. Selain itu, data yang digunakan hanya terbatas pada tweet berbahasa Indonesia, yang mungkin tidak mencakup seluruh spektrum opini publik di Indonesia. Untuk penelitian mendatang, penting untuk memperluas sumber data, misalnya dengan menggunakan data dari platform lain yang dapat memberikan pandangan lebih luas mengenai sentimen publik terhadap Starlink. Secara keseluruhan, penelitian ini memberikan pandangan berharga bagi Starlink dalam memahami pandangan publik terhadap layanan mereka di Indonesia. Dengan memahami sentimen yang ada. Selain itu, penelitian ini membuka peluang untuk pengembangan lebih lanjut dalam teknik analisis sentimen yang lebih akurat, dengan penerapan model pembelajaran mesin yang lebih canggih di masa depan.

4. KESIMPULAN

Hasil analisis sentimen terhadap layanan internet Starlink di Indonesia dengan menggunakan algoritma Naïve Bayes dengan teknik TF-IDF dan *Cross-Validation* menunjukkan bahwa sentimen positif mendominasi dengan 49,38%, terkait dengan kecepatan dan stabilitas layanan. Sentimen netral mencapai 32,94%, yang mencerminkan fokus percakapan pada aspek teknis tanpa kecenderungan emosional, sementara sentimen negatif tercatat 17,68%, lebih banyak terkait dengan harga tinggi dan kekhawatiran terhadap dampak terhadap industri lokal. Secara keseluruhan, sentimen positif lebih dominan, disebabkan oleh apresiasi publik terhadap kecepatan dan stabilitas layanan Starlink, yang dianggap sebagai solusi untuk keterbatasan konektivitas di daerah terpencil. Penting bagi Starlink untuk mempertahankan kualitas layanan, terutama kecepatan dan stabilitas, sambil mempertimbangkan penyesuaian harga untuk mengatasi sentimen negatif. Metode yang digunakan terbukti efektif dalam mengklasifikasikan sentimen, dengan TF-IDF memberikan bobot pada kata signifikan dan *Cross-Validation* memastikan generalisasi yang baik.

Keterbatasan penelitian ini terletak pada ketidakseimbangan data, dengan sentimen positif yang lebih dominan, yang mempengaruhi klasifikasi sentimen negatif dan netral. Model ini juga terbatas pada data tweet berbahasa Indonesia, yang dapat mempengaruhi keberagaman opini publik. Penelitian mendatang perlu

mengembangkan model dengan teknik yang lebih canggih, seperti deep learning, serta memperluas sumber data.

REFERENSI

[1] W. Rakhmawati, C. E. Kosasih, R. Widiasih, S. Suryani, and H. Arifin, "Internet Addiction Among Male Adolescents in Indonesia: A Qualitative Study," *Am. J. Mens. Health*, vol. 15, no. 3, 2021, doi: 10.1177/15579883211029459.

[2] J. Caron and J. R. Markusen, "Importancy Of Internet," vol. 2, no. 13, pp. 1–23, 2016.

[3] APJII, "APJII Jumlah Pengguna Internet Indonesia Tembus 221 Juta Orang," APJII. Accessed: Aug. 07, 2024. [Online]. Available: https://apjii.or.id/berita/d/apjii-jumlah-pengguna-internet-indonesia-tembus-221-juta-orang

[4] R. Dewantara, P. A. Cakranegara, A. J. Wahidin, A. Muditomo, I. Gede, and I. Sudipa, "Implementasi Metode Preference Selection Index Dalam Penentuan Jaringan Dan Pemanfaatan Internet Pada Provinsi Indonesia," *J. Sains Komput. Inform. (J-SAKTI*, vol. 6, no. 2, pp. 1226–1238, 2022.

[5] ASSI, "Starlink Masuk Indonesia, Pengusaha Lokal Mulai Tingkatkan Kapasitas Satelit," ASSI. [Online]. Available: https://apsat.assi.or.id/2024/06/07/starlink-masuk-indonesia-pengusaha-lokal-mulai-tingkatkan-kapasitas-satelit/#:~:text=Sebelumnya Elon Musk meresmikan peluncuran,19%2F05%2F2024).

[6] T. Duan and V. Dinavahi, "Starlink Space Network-Enhanced Cyber–Physical Power System," *IEEE Trans. Smart Grid*, vol. 12, no. 4, pp. 3673–3675, 2021, doi: 10.1109/TSG.2021.3068046.

[7] M. S. Kalaivani, S. Jayalakshmi, and R. Priya, "Comparative analysis of sentiment classification using machine learning techniques on Twitter data," *Int. J. Health Sci. (Qassim).*, vol. 6, no. April, pp. 8273–8280, 2022, doi: 10.53730/ijhs.v6ns2.7098.

[8] M. A. Saddam, E. K. Dewantara, and A. Solichin, "Sentiment Analysis of Flood Disaster Management in Jakarta on Twitter Using Support Vector Machines," *Sinkron*, vol. 8, no. 1, pp. 470–479, 2023, doi: 10.33395/sinkron.v8i1.12063.

[9] B. Liu, *Sentiment analysis: Mining opinions, sentiments, and emotions*. 2015. doi: 10.1017/CBO9781139084789.

[10] M. Wankhade, A. C. S. Rao, and C. Kulkarni, *A survey on sentiment analysis methods, applications, and challenges*, vol. 55, no. 7. Springer Netherlands, 2022. doi: 10.1007/s10462-022-10144-1.

[11] wesam ahmed, N. Semary, K. Amin, and M. Adel Hammad, "Sentiment Analysis on Twitter Using Machine Learning Techniques and TF-IDF Feature Extraction: A Comparative Study," *IJCI. Int. J. Comput. Inf.*, vol. 10, no. 3, pp. 52–57, 2023, doi: 10.21608/ijci.2023.236052.1128.

[12] R. O. M. Khalil Gibran1, Mhd Ikhsan Rifki2, Abdul Halim Hasugian3,Ahmad Taufik Al Afkari Siahaan4, Afandi Sahputra5, "Sentiment Analysis of Platform X Users on Starlink Using Naive Bayes," *Instal J. Komput.*, vol. 10, no. July, 2024, [Online]. Available: https://journalinstal.cattleyadf.org/index.php/Instal/article/view/240

[13] S. Sardin, A. Nugroho, and N. T. Kurniadi, "Sentiment Analysis of Starlink on Twitter Using Support Vector Machine Algorithm," *J. Comput. Networks, Archit. High Perform. Comput.*, vol. 6, no. 3, pp. 1321–1332, 2024, doi: 10.47709/cnahpc.v6i3.4348.

[14] N. Nicholas and R. Sutomo, "Comparative Analysis of Sentiment Analysis Using the Support Vector Machine and Naive Bayes Algorithm on Cryptocurrencies," *J. Multidiscip. Issues*, vol. 1, no. 3, pp. 2–19, 2021, doi: 10.53748/jmis.v1i3.22.

[15] S. M. Permataning Tyas, B. S. Rintyarna, and W. Suharso, "The Impact of Feature Extraction to Naïve Bayes Based Sentiment Analysis on Review Dataset of Indihome Services," *Digit. Zo. J. Teknol. Inf. dan Komun.*, vol. 13, no. 1, pp. 1–10, 2022, doi: 10.31849/digitalzone.v13i1.9158.

[16] A. Rahmatulloh, R. N. Shofa, I. Darmawan, and Ardiansah, "Sentiment Analysis of Ojek Online User Satisfaction Based on the Naïve Bayes and Net Brand Reputation Method," in *2021 9th International Conference on Information and Communication Technology (ICoICT)*, 2021, pp. 337–341. doi: 10.1109/ICoICT52021.2021.9527466.

[17] A. Addiga and S. Bagui, "Sentiment Analysis on Twitter Data Using Term Frequency-Inverse Document Frequency," *J. Comput. Commun.*, vol. 10, no. 08, pp. 117–128, 2022, doi: 10.4236/jcc.2022.108008.

[18] Y. Zhang, Y. Zhou, and J. T. Yao, *Feature Extraction with TF-IDF and Game-Theoretic Shadowed Sets*, vol. 1237 CCIS. Springer International Publishing, 2020. doi: 10.1007/978-3-030-50146-4_53.

[19] Lisnawati, "Kehadiran Starlink di Indonesia : Manfaat dan Dampak," *Info Singk. Kaji. Singk. Tehadap isu Aktual dan Strateg.*, vol. 16, no. 11, pp. 16–20, 2024.

[20] BBC, "Pro-kontra Starlink di Indonesia - 'Pemain lokal juga mampu, pemerintah jangan anak emaskan pemain asing,'" BBC. [Online]. Available: https://www.bbc.com/indonesia/articles/cmll91z484ro

[21] H. R. Alhakiem and E. B. Setiawan, "Aspect-Bas1ed Sentiment Analysis on Twitter Using Logistic

Regression with FastText Feature Expansion," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, 2022, [Online]. Available: https://api.semanticscholar.org/CorpusID:253335745

[22] I. S. Thalib, S. K. Gusti, F. Yanto, and M. Affandes, "Klasifikasi Sentimen Tragedi Kanjuruhan Pada Twitter Menggunakan Algoritma Naïve Bayes," *J. Sist. Komput. dan Inform.*, vol. 4, no. 3, p. 467, 2023, doi: 10.30865/json.v4i3.5852.

[23] E. Puspita, D. F. Shiddieq, and F. F. Roji, "Pemodelan Topik pada Media Berita Online Menggunakan Latent Dirichlet Allocation (Studi Kasus Merek Somethinc)," *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 4, no. 2, pp. 481–489, 2024, doi: 10.57152/malcom.v4i2.1204.

[24] I. Verawati and S. N. Jaelani, "JURNAL MEDIA INFORMATIKA BUDIDARMA Analisis Sentimen Pengguna Twitter Terhadap Bus Listrik Menggunakan Naïve Bayes," *J. Media Inform. Budidarma*, vol. 8, no. 2, pp. 832–842, 2024, doi: 10.30865/mib.v8i2.7030.

[25] L. Xiang, "Application of an Improved TF-IDF Method in Literary Text Classification," *Adv. Multimed.*, vol. 2022, 2022, doi: 10.1155/2022/9285324.

[26] R. Kosasih and A. Alberto, "Sentiment analysis of game product on shopee using the TF-IDF method and naive bayes classifier," *Ilk. J. Ilm.*, vol. 13, no. 2, pp. 101–109, 2021, doi: 10.33096/ilkom.v13i2.721.101-109.

[27] Y. D. Kirana and S. Al Faraby, "Sentiment analysis of beauty product reviews using the K-nearest neighbor (KNN) and TF-IDF methods with chi-square feature selection," 2021, *scholar.archive.org*. [Online]. Available: https://scholar.archive.org/work/ye6ofgjo45ey5fahm5jmjr55za/access/wayback/https://commdis.telkomuniversity.ac.id/jdsa/index.php/jdsa/article/download/71/31/

[28] T. N. Viet *et al.*, "The Naïve Bayes Algorithm," vol. 12, no. 4, pp. 1038–1043.

[29] G. I. Webb, "Notsonaive," *Not So Naive Bayes Aggregating One-Dependence Estim.*, vol. 58, no. 5–24, pp. 5–24, 2005, [Online]. Available: papers://5e3e5e59-48a2-47c1-b6b1-a778137d3ec1/Paper/p1997

[30] T. Ridwansyah, "Implementasi Text Mining Terhadap Analisis Sentimen Masyarakat Dunia Di Twitter Terhadap Kota Medan Menggunakan K-Fold Cross Validation Dan Naïve Bayes Classifier," *KLIK Kaji. Ilm. Inform. dan Komput.*, vol. 2, no. 5, pp. 178–185, 2022, doi: 10.30865/klik.v2i5.362.

[31] A. Seraj *et al.*, "Chapter 5 - Cross-validation," S. Eslamian and F. B. T.-H. of H. Eslamian, Eds., Elsevier, 2023, pp. 89–105. doi: https://doi.org/10.1016/B978-0-12-821285-1.00021-X.

[32] R. S. Kharisma, Muttafi'ah, and A. Dahlan, "Comparison of Naïve Bayes Algorithm Model Combinations with Term Weighting Techniques in Sentiment Analysis," in *2021 4th International Conference on Information and Communications Technology (ICOIACT)*, 2021, pp. 160–163. doi: 10.1109/ICOIACT53268.2021.9563999.

[33] H. Kaur and N. K. Sandhu, "International Journal of Communication Networks and Information Security Evaluating the Effectiveness of the Proposed System Using F1 Score , Recall , Accuracy , Precision and Loss Metrics Compared to Prior Techniques," vol. 15, no. 04, pp. 368–383, 2023.