ANALISIS SENTIMEN STARLINK DI INDONESIA DENGAN PENDEKATAN LEXICON DAN RANDOM FOREST



Disusun Oleh:

N a m a : Ivan Aditya Nugraha

NIM : 20523080

PROGRAM STUDI INFORMATIKA – PROGRAM SARJANA

FAKULTAS TEKNOLOGI INDUSTRI

UNIVERSITAS ISLAM INDONESIA

2025

HALAMAN PENGESAHAN DOSEN PEMBIMBING

ANALISIS SENTIMEN STARLINK DI INDONESIA DENGAN PENDEKATAN LEXICON DAN RANDOM FOREST

TUGAS AKHIR

Disusun Oleh:

N a m a : Ivan Aditya Nugraha

NIM : 20523080

Yogyakarta, 14 Juli 2025

Pembimbing,

(Irving Vitra Paputungan, S.T., M.SC., PH.D.)

HALAMAN PENGESAHAN DOSEN PENGUJI

ANALISIS SENTIMEN STARLINK DI INDONESIA DENGAN PENDEKATAN LEXICON DAN RANDOM FOREST

TUGAS AKHIR

Telah dipertahankan di depan sidang penguji sebagai salah satu syarat untuk memperoleh gelar Sarjana Komputer dari Program Studi Informatika – Program Sarjana di Fakultas Teknologi Industri Universitas Islam Indonesia Yogyakarta, 4 Agustus 2025

Tim Penguji

Irving Vitra Paputungan, S.T., M.Sc., Ph.D. _____

Anggota 1
Dhomas Hatta Fudholi, S.T., M.Eng., Ph.D. _____

Anggota 2

Elyza Gustri Wahyuni, S.T., M.Cs. _____

Mengetahui,

Ketua Program Studi Informatika – Program Sarjana Fakultas Teknologi Industri
Universitas Islam Indonesia

(Dhomas Hatta Fudholi, S.T., M.Eng., Ph.D.)

HALAMAN PERNYATAAN KEASLIAN TUGAS AKHIR

Yang bertanda tangan di bawah ini:

Nama : Ivan Aditya Nugraha

NIM : 20523080

Tugas akhir dengan judul:

ANALISIS SENTIMEN STARLINK DI INDONESIA DENGAN PENDEKATAN LEXICON DAN RANDOM FOREST

Menyatakan bahwa seluruh komponen dan isi dalam tugas akhir ini adalah hasil karya saya sendiri. Apabila di kemudian hari terbukti ada beberapa bagian dari karya ini adalah bukan hasil karya sendiri, tugas akhir yang diajukan sebagai hasil karya sendiri ini siap ditarik kembali dan siap menanggung risiko dan konsekuensi apapun.

Demikian surat pernyataan ini dibuat, semoga dapat dipergunakan sebagaimana mestinya.

Yogyakarta, 4 Agustus 2025



(Ivan Aditya Nugraha)

HALAMAN PERSEMBAHAN

Alhamdulillahirobbil ‘alamin, puji syukur saya kepada Allah SWT, berkat karuniaNya, penulis dapat menyelesaikan tugas akhir ini. Maka dari itu, saya persembahkan tugas akhir saya kepada:

Kedua orang tua saya, yaitu Soemarsono dan Frida Heny Kusumawanti yang selalu memberikan dukungan melalui doa – doa, semangat, dan motivasi selama proses pengerjaan tugas akhir ini.

Adik saya yaitu Annisa Aulia Devitasari yang baru saja memasuki masa perkuliahan prodi Ilmu Komunikasi agar selalu semangat dalam menjalani masa perkuliahan.

Dosen pembimbing saya, yaitu Irving Vitra Paputungan, S.T., M.Sc., Ph.D., atas arahan selama masa pengerjaan tugas akhir ini.

Kemudian yang terakhir kepada teman – teman dekat yang telah menemani serta saling *support* selama proses pengerjaan tugas akhir.

HALAMAN MOTO

"Failure is an option here. If things are not failing, you are not innovating enough."

Elon Musk

"It's not the future I'm afraid of. It's repeating the past."

Kurisu Makise, Steins;Gate

KATA PENGANTAR

Assalamualaikum Warahmatullahi Wabarakatuh.

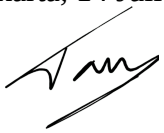
Alhamdulillahirobbil ‘alamin, puji syukur saya kepada Allah SWT atas karunia-Nya, sehingga penulis dapat menyelesaikan penyusunan tugas akhir yang berjudul “Analisis Sentimen Starlink di Indonesia dengan Pendekatan Lexicon dan Random Forest” sebagai salah satu syarat untuk memperoleh gelar Sarjana Komputer di Program Studi Informatika – Fakultas Teknologi Industri, Universitas Islam Indonesia.

Penulis ingin berterimakasih atas dukungan serta doa dari banyak pihak selama pengerjaan tugas akhir ini. Penulis ingin menyampaikan ucapan terima kasih yang sebesar – besarnya kepada:

1. Tuhan Yang Maha Esa, Allah Subhanallahu wa ta’ala atas segala Rahmat yang diberikan-Nya sehingga penulis dapat menyelesaikan tugas akhir ini dengan lancar.
2. Kedua orang tua serta keluagra besar yang selalu memberikan *support* melalui doa dan material yang sangat membantu penulis dalam menyelesaikan tugas akhir ini.
3. Bapak Irving Vitra Paputungan, S.T., M.Sc., Ph.D., selaku dosen pembimbing penulis yang senantiasa memberi bimbingan dalam penyusunan tugas akhir ini.
4. Seluruh dosen dan staf Jurusan Program Studi Informatika – Program Sarjana Universitas Islam Indonesia atas ilmu yang telah diberikan selama masa studi.
5. Seluruh teman – teman di grup “Tua Renta Tak Berdaya”, Dawwam, Eki, Faris, Hudha, Raihan, Rama, Reza, Ripal, Sulthon, Thoriq yang telah menjadi teman dekat di masa perkuliahan dan saling *support* terkait tugas akhir.
6. Seluruh teman – teman di grup “WIBU – Season Finale” yang telah banyak sekali membantu selama masa perkuliahan dan saling mendukung terkait tugas akhir.
7. Terima kasih kepada Anjeli atas bantuan dan semangatnya yang telah mendukung penulis dalam menyelesaikan tugas akhir.

Semoga tugas akhir ini dapat bermanfaat bagi orang lain. Akhir kata penulis ucapkan mohon maaf atas segala kekurangan dan terima kasih.

Yogyakarta, 14 Juli 2025



(Ivan Aditya Nugraha)

SARI

Perkembangan internet yang pesat mendorong kebutuhan akan koneksi stabil di seluruh wilayah Indonesia. Starlink hadir sebagai layanan internet berbasis satelit orbit rendah (LEO) yang menjanjikan konektivitas di area terpencil. Penelitian ini bertujuan untuk menganalisis persepsi publik terhadap layanan Starlink berdasarkan komentar pengguna YouTube, dengan menggunakan pendekatan analisis sentimen berbasis lexicon dan algoritma Random Forest.

Data dikumpulkan dari dua video YouTube populer tentang Starlink, kemudian dilakukan preprocessing yang mencakup *cleaning*, normalisasi, tokenisasi, dan *stemming*. Sentimen diklasifikasikan menjadi tiga, yaitu positif, negatif, dan netral menggunakan *lexicon* Bahasa Indonesia. Selanjutnya, model klasifikasi dikembangkan menggunakan algoritma Random Forest dan diuji pada tiga skenario pembagian data: 60:40, 70:30, dan 80:20. Evaluasi dilakukan berdasarkan akurasi, presisi, recall, dan F1-score.

Hasil penelitian menunjukkan bahwa model mampu memprediksi sentimen dengan performa terbaik pada rasio 80:20. Penelitian ini juga menghasilkan model prediksi sederhana untuk mengklasifikasikan komentar baru secara otomatis.

Kata kunci: Analisis sentiment, komentar *YouTube*, *lexicon*, *random forest*, Starlink

GLOSARIUM

Analisis Sentimen	Proses identifikasi dan ekstraksi opini atau emosi dari teks.
Dataset	Kumpulan data yang digunakan untuk pelatihan dan pengujian model.
Lexicon	Kamus kata yang digunakan untuk menentukan nilai sentimen positif atau negatif dari suatu kata.
Random Forest	Algoritma *machine learning* berbasis *ensemble* yang menggabungkan banyak *decision tree* untuk klasifikasi.
Sastrawi	*Library Python* untuk pemrosesan teks bahasa Indonesia, khususnya *stemming* dan *stopword removal.*

DAFTAR ISI

HALAMAN JUDUL i
HALAMAN PENGESAHAN DOSEN PEMBIMBING ii
HALAMAN PENGESAHAN DOSEN PENGUJI iii
HALAMAN PERNYATAAN KEASLIAN TUGAS AKHIR iv
HALAMAN PERSEMBAHAN v
HALAMAN MOTO vi
KATA PENGANTAR vii
SARI viii
GLOSARIUM ix
DAFTAR ISI x
DAFTAR TABEL xii
DAFTAR GAMBAR xiii
BAB I PENDAHULUAN 1
1.1 Latar Belakang 1
1.2 Rumusan Masalah 6
1.3 Batasan Masalah 6
1.4 Tujuan Penelitian 7
1.5 Manfaat Penelitian 7
1.6 Metode Penelitian 8
1.7 Sistematika Penelitian 8
BAB II LANDASAN TEORI DAN KAJIAN PUSTAKA 10
2.1 Landasan Teori 10
2.1.1 Analisis Sentimen 10
2.1.2 Klasifikasi Teks 10
2.1.3 Natural Language Processing (NLP) 10
2.1.4 Text Mining 11
2.1.5 Web Scraping 11
2.1.6 Random Forest 11
2.1.7 Lexicon-Based Approach 12
2.2 Kajian Pustaka 12
BAB III METODOLOGI PENELITIAN 17
3.1 *Data Collection* (Pengumpulan Data) 18
3.2 Preprocessing Data 20

3.3 Processing Data 29
3.4 Rencana Pengujian 34
BAB IV HASIL DAN PEMBAHASAN 36
4.1 Pengujian Skenario 36
4.1.1 Pembagian Data 36
4.1.2 Algoritma 38
4.1.3 Evaluasi Performa Model 39
4.1.4 Modelling 48
4.1.5 Testing 49
BAB V KESIMPULAN DAN SARAN 54
5.1 Kesimpulan 54
5.2 Saran 55
5.3 Penutup 56
DAFTAR PUSTAKA 57
LAMPIRAN 62

DAFTAR TABEL

Tabel 2.1 Kajian Pustaka 12
Tabel 3.1 *Lexicon Dictionary* Bahasa Indonesia Sentimen Positif 30
Tabel 3.2 *Lexicon Dictionary* Bahasa Indonesia Sentimen Negative 30
Tabel 3.3 Total Data Sentimen 31
Tabel 4.1 Performa Ketiga Model 47
Tabel 4.2 Daftar Kalimat untuk Pengujian Model 49
Tabel 4.3 Daftar Kalimat dan Hasil Pengujian Model 51

DAFTAR GAMBAR

Gambar 1.1 Elon Musk *(@elonmusk)* Mengumumkan Peluncuran Starlink 2
Gambar 1.2 NanangMrk Review Starlink 3
Gambar 1.3 GadgetIn Review Starlink 4
Gambar 3.1 Diagram Alur Penelitian 18
Gambar 3.2 Dataset Top Comments dari Kanal GadgetIn 19
Gambar 3.3 Dataset Newest Comments dari Kanal GadgetIn 19
Gambar 3.4 Dataset Top Comments dari Kanal NanangMrk 19
Gambar 3.5 Dataset Newest Comments dari Kanal NanangMrk 19
Gambar 3.6 Dataset Gabungan dari Kedua Video 20
Gambar 3. 7 Total Komentar Terduplikasi 20
Gambar 3.8 Tahapan Prepocessing Data 21
Gambar 3.9 *Delete Duplicates* 22
Gambar 3.10 *Diagram Alur Text Cleaning* 22
Gambar 3.11 Sebelum dibersihkan dan *lowercasing* 25
Gambar 3.12 Sesudah dibersihkan dan *lowercasing* 25
Gambar 3.13 Sebelum Normalisasi 26
Gambar 3.14 Sesudah Normalisasi 26
Gambar 3.15 Sebelum *Tokenizing* 27
Gambar 3.16 Sesudah *Tokenizing* 27
Gambar 3.17 Sebelum *Stopword Removal* 28
Gambar 3.18 Sesudah *Stopword Removal* 28
Gambar 3.19 Sebelum *Stemming* 29
Gambar 3.20 Sesudah *Stemming* 29
Gambar 3.21 Dataset Setelah Diberi Label Sentimen 31
Gambar 3.22 Visualisasi Kata yang Bernilai Positif dalam Bentuk *Wordcloud* 32
Gambar 3.23 Visualisasi Kata yang Bernilai Sentimen Negatif dalam Bentuk *Wordcloud* 33
Gambar 3.24 Visualisasi Kata yang Bernilai Sentimen Netral dalam Bentuk *Wordcloud* 34
Gambar 4.1 Import Library untuk Pembagian Data 36
Gambar 4.2 Implementasi Kode Python untuk Pembagian Data Latih dan Data Uji dengan Rasio 60:40 37
Gambar 4.3 Implementasi Kode Python untuk Pembagian Data Latih dan Data Uji dengan Rasio 70:30 37

Gambar 4.4 Implementasi Kode Python untuk Pembagian Data Latih dan Data Uji dengan Rasio 80:20 38
Gambar 4.5 Implementasi Model Random Forest Menggunakan Pipeline dari Scikit-Learn. 38
Gambar 4.6 Confusion Matrix 39
Gambar 4.7 Hasil Evaluasi Model pada Rasio 60:40 40
Gambar 4.8 Confusion Matrix pada Rasio 60:40 41
Gambar 4.9 Hasil Evaluasi Model pada Rasio 70:30 43
Gambar 4.10 Confusion Matrix pada Rasio 70:30 43
Gambar 4.11 Hasil Evaluasi Model pada Rasio 80:20 45
Gambar 4.12 Confusion Matrix pada Rasio 80:20 46
Gambar 4.13 Model Prediksi Sentimen 49
Gambar 4.14 Output dari Model Prediksi Sentimen 49

BAB I
PENDAHULUAN

1.1 Latar Belakang

Di era digital saat ini, perkembangan teknologi informasi dan komunikasi mengalami peningkatan yang luar biasa, mengubah hampir seluruh aspek kehidupan manusia. Salah satu perubahan terbesar adalah peningkatan penggunaan internet dan media sosial sebagai sarana komunikasi, informasi, dan ekspresi diri. Menurut (Rahman & Munawaroh, n.d.), berdasarkan hasil survei Asosiasi Penyelenggara Jasa Internet di Indonesia (APJII), pengguna internet di Indonesia mencapai 215,63 juta jiwa pada periode tahun 2022 – 2023. Angka ini menunjukkan betapa pentingnya peran internet dan media sosial dalam kehidupan sehari – hari masyarakat Indonesia.

Internet merupakan sebuah hal penting terhadap perkembangan era digital saat ini (Nisa et al., 2024). Konektivitas internet yang dapat diandalkan telah menjadi suatu kewajiban di era digital saat ini. Salah satu infrastruktur paling krusial yang mendukung perkembangan teknologi ini adalah jaringan internet (Putra et al., 2024). Jaringan internet yang stabil dan cepat memungkinkan pengguna untuk tetap produktif dan terhubung di era digital ini. Perkembangan teknologi yang pesat turut mendorong laju pertumbuhan internet, di mana meningkatnya kebutuhan akan informasi dan pengetahuan menjadi salah satu pendorong utamanya (Akmala, 2019).

Seiring waktu, internet juga mengalami evolusi yang signifikan. Awalnya, internet hanya dapat diakses melalui jaringan kabel yang terbatas di area tertentu. Namun, perkembangan teknologi nirkabel, seperti Wi-Fi dan 4G, telah membuat internet menjadi lebih mudah diakses di mana saja dan kapan saja. Inovasi ini terus berlanjut dengan pengembangan jaringan 5G yang menawarkan kecepatan lebih tinggi dan latensi lebih rendah.

Salah satu terobosan terbaru dalam teknologi internet adalah Starlink. Starlink merupakan layanan internet yang berbasis *Low Earth Orbit (LEO)*, dalam kata lain adalah satelit orbit rendah yang dikembangkan oleh SpaceX, perusahaan yang didirikan Elon Musk ((Lisnawati, 2024). Starlink bertujuan untuk menyediakan akses internet global melalui armada satelit yang terbang sekitar 500km di atas permukaan bumi ((Mohan et al., 2024).

Meskipun Starlink menjadi salah satu pemain paling menonjol, industri internet satelit LEO saat ini sedang mengalami perkembangan pesat dan memunculkan beberapa pesaing global. Di antaranya adalah Project Kuiper, inisiatif layanan internet satelit dari Amazon, dan

OneWeb, sebuah konstelasi satelit LEO yang berfokus pada pasar bisnis dan pemerintah. Konstelasi OneWeb ini diusulkan oleh perusahaan Inggris, OneWeb, yang sebelumnya bernama WorldVu, untuk menyediakan akses internet global kepada konsumen individu dan maskapai penerbangan (Zhi et al., 2024). Baik Project Kuiper maupun OneWeb memiliki tujuan yang sama dengan Starlink, yaitu menghadirkan konektivitas berkecepatan tinggi ke wilayah yang sulit dijangkau. Keberadaan para pemain ini menunjukkan bahwa teknologi internet satelit LEO telah menjadi sektor yang sangat kompetitif secara global, menawarkan solusi alternatif bagi tantangan konektivitas di berbagai belahan dunia.

Di Indonesia pada 19 Mei, Elon Musk secara resmi mengumumkan peluncuran Starlink di Indonesia melalui sebuah unggahan akun pribadinya di platform X (sebelumnya Twitter) seperti pada Gambar 1.1. Dalam unggahan tersebut, Elon Musk menyatakan rasa hormatnya atas peluncuran ini, langkah tersebut menandai era baru akses internet di Indonesia, terutama bagi wilayah yang selama ini sulit dijangkau oleh infrastruktur tradisional. Peluncuran ini disambut antusias oleh warganet Indonesia, yang berharap Starlink dapat menjadi solusi efektif untuk memperbaiki kualitas dan aksesibilitas internet, terutama di daerah pedalaman yang selama ini mengalami keterbatasan koneksi. Antusiasme ini tercermin dari puluhan ribu retweet dan ratusan ribu tanda suka pada unggahan tersebut.



Gambar 1.1 Elon Musk *(@elonmusk)* Mengumumkan Peluncuran Starlink

Peluncuran *Starlink* telah menarik perhatian publik, terutama di kalangan pengguna internet dan penggemar teknologi. Banyak *YouTuber* dan *reviewer* teknologi yang telah mencoba layanan ini dan membagikan pengalaman mereka di platform video tersebut. Ulasan – ulasan ini memberikan gambaran tentang performa *Starlink*, mulai dari kecepatan *download*

dan *upload*, stabilitas koneksi, hingga kemudahan instalasi perangkat. Beberapa ulasan juga menyoroti tantangan yang dihadapi, seperti gangguan cuaca yang dapat mempengaruhi koneksi, serta biaya langganan yang relatif tinggi.

Kanal *YouTube* pertama yang membahas tentang *Starlink* di Indonesia adalah NanangMrk, dengan videonya yang berjudul "STARLINK INDONESIA REVIEW, INTERNET DARI LANGIT TANPA SINYAL DIMANAPUN ANDA BERADA (TERLENGKAP)" seperti pada Gambar 1.2. Video ini diunggah di *YouTube* pada 11 Mei 2024. Dalam video ini, menampilkan proses *unboxing* perangkat *Starlink*, konfigurasi awal, serta pengujian performa layanan internet satelit tersebut. Nanang, sebagai kreator sekaligus pemilik kanal, menjelaskan berbagai aspek *Starlink* mulai dari spesifikasi perangkat keras, kecepatan *download* dan *upload*, hingga kelebihan dan tantangan yang dihadapi saat menggunakan layanan Starlink di Indonesia.



Gambar 1.2 NanangMrk Review Starlink

Video ini mendapatkan respons yang besar dari publik, terbukti dengan jumlah penayangan yang mencapai 1,02 juta *views* dan mendapatkan 4.837 komentar. Berbagai komentar yang dilontakran oleh penonton mencakup pujian, kritik, pertanyaan teknis, serta harapan terhadap pengembangan layanan Starlink di masa depan.

Tidak hanya NanangMrk yang membahas tentang *Starlink* di Indonesia, salah satu kanal *YouTube* teknologi populer, GadgetIn, juga membahas tentang layanan *Starlink* melalui video berjudul "Review Starlink – Ini Revolusioner". Video ini diunggah enam bulan lalu dan telah ditonton sebanyak 3,6 juta kali serta mendapat 89 ribu *like* dari penonton. David Brendi, selaku pemilik kanal YouTube GadgetIn menguji kemampuan Starlink di daerah terperincil yang

dapat dilihat pada Gambar 1.3. Dalam video berdurasi 24 menit ini, pemilik kanal GadgetIn memberikan ulasan mengenai performa Starlink, termasuk kecepatan internet, kestabilan koneksi, dan proses instalasi perangkat. Video ini direkam di lokasi terpencil untuk menunjukkan kemampuan Starlink dalam menyediakan koneksi internet di daerah yang sulit dijangkau jaringan konvensional.



Gambar 1.3 GadgetIn Review Starlink

Seiring dengan meningkatnya atensi publik terhadap layanan Starlink di Indonesia, terutama melalui berbagai konten ulasan di *YouTube*, muncul banyak sekali respon dan opini dari masyarakat yang terlihat pada kolom komentar. Maka, diperlukan pendekatan analisis yang mampu mengkaji presepsi publik dalam bentuk teks. Dalam hal ini, analisis sentimen menjadi salah satu metode yang tepat untuk digunakan. Menurut (Fatma Sjoraida et al., 2024), analisis sentimen merupakan pendekatan yang digunakan untuk mengevaluasi opini dalam teks, dan telah menjadi fokus utama dalam memahami manusia merespons berbagai jenis konten.

Terdapat tujuh alasan utama mengapa analisis sentimen penting. Pertama, karena mampu menggali wawasan dari data tidak terstruktur, seperti ribuan komentar yang tidak bisa dianalisis secara manual. (Mata, n.d.) menyatakan bahwa pertumbuhan opini di media sosial dan forum mendorong kebutuhan analisis sentimen, karena opini dapat mempengaruhi perilaku, kepercayaan, dan persepsi manusia dalam berbagai aspek kehidupan. Kedua, analisis sentimen mendukung pengambilan keputusan bisnis, terutama dalam memahami kebutuhan dan harapan

konsumen. (Kauffmann et al., 2019) menjelaskan bahwa *user-generated content* (UGC) merupakan salah satu sumber utama *big data* yang bernilai strategis dalam pemasaran, karena informasi dari konsumen memiliki potensi besar dalam mendukung keunggulan kompetitif perusahaan. Ketiga, analisis sentiment dapat membantu dalam memahami persepsi publik terhadap produk atau layanan baru, seperti Starlink. Menurut (Kauffmann et al., 2019), analisis sentimen dapat digunakan untuk mengelompokkan opini pelanggan, memahami motivasi di balik preferensi mereka terhadap suatu produk atau layanan, serta memprediksi perilaku konsumen berdasarkan umpan balik yang diperoleh dari *platform* digital. Keempat, analisis ini berfungsi sebagai alat *monitoring* reputasi dan manajemen *brand*. Menurut (Aftab et al., 2023), analisis sentimen memberi wawasan tentang sikap dan persepsi pelanggan terhadap merek, memungkinkan bisnis untuk mengidentifikasi sentimen negatif dan menyusun strategi brand yang lebih terfokus sebelum reputasi terdampak. Kelima, analisis sentimen berperan dalam meningkatkan pengalaman pelanggan. (Karne, 2022) menyebutkan bahwa analisis sentimen digunakan untuk memahami faktor emosional yang mempengaruhi perilaku pelanggan, seperti rasa senang, frustrasi, atau kepercayaan, yang pada gilirannya memengaruhi keputusan pembelian, loyalitas, dan kepuasan pelanggan secara keseluruhan. Keenam, analisis sentimen mendukung pengembangan model prediktif yang berguna dalam berbagai domain. (Cortis & Davis, 2021) mencatat bahwa *opinion mining* dari media sosial telah digunakan secara luas dalam pemasaran, periklanan, penjualan, serta membantu pengembangan sistem prediksi seperti layanan pelanggan dan pengambilan keputusan berdasarkan pola opini publik. Dan terakhir, analisis sentimen memberikan efisiensi dan otomatisasi dalam memproses opini publik dalam jumlah besar. Menurut (Mata & Martins, n.d.), analisis sentimen memungkinkan komputer untuk mengotomatiskan aktivitas yang biasanya dilakukan manusia dalam pengambilan keputusan berdasarkan opini, dengan aplikasi luas dalam *data mining* dan *web mining*.

Untuk melakukan analisis sentimen, salah satu pendekatan yang umum digunakan adalah *lexicon-based sentiment analysis* atau analisis sentimen berbasis leksikon. Pendekatan berbasis leksikon digunakan untuk memvisualisasikan dan menganalisis emosi pada kumpulan data yang sama. Pendekatan berbasis *lexicon* menghitung sentimen sebuah dokumen dengan mengidentifikasi nilai positif atau negatif pada setiap kata di dalamnya. Proses ini dilakukan dengan mencocokkan kata – kata tersebut dengan kamus leksikon. Pendekatan berbasis leksikon adalah metode yang sangat efektif dan mudah (Ainapure et al., 2023).

Selain itu, diperlukan pemanfaatan algoritma *machine learning.* Algoritma yang digunakan adalah *Random Forest.* Algoritma ini dipilih karena merupakan algoritma *machine learning* yang mudah digunakan serta merupakan salah satu algoritma yang sering digunakan karena dapat digunakan sebagai klasifikasi dan regresi (Azhari et al., 2021). Model *Random Forest* juga merupakan algoritma *ensemble* yang memprediksi sentimen dengan mengintegrasikan hasil dari serangkaian pohon keputusan. Nama "*Random*" merujuk pada proses di mana setiap pohon dibangun dari subset data dan fitur yang dipilih secara acak. Pendekatan ini secara signifikan mengurangi risiko *overfitting*, sehingga menghasilkan model dengan akurasi dan stabilitas yang lebih tinggi (Ramadhansyah et al., 2024). Selain itu, seleksi fitur yang efektif berperan penting dalam mengidentifikasi variabel prediktif paling relevan. Hal ini secara signifikan dapat meningkatkan akurasi dan efisiensi model klasifikasi (Herjanto & Carudin, 2024). Oleh karena itu, penelitian ini juga menguji kemampuan algoritma *Random Forest* dalam mengklasifikasikan sentimen berdasarkan komentar – komentar pengguna di *YouTube.*

Dengan menganalisis data komentar dari kedua video tersebut, penelitian ini tidak hanya meliputi menggali informasi opini publik, namun juga menghitung proporsi komentar positif, negatif, dan netral, mengevaluasi hasil analisis sentimen dengan algoritma *Random Forest*, serta menilai kemampuan model dalam memprediksi sentimen dari kalimat baru.

1.2 Rumusan Masalah

Berdasarkan latar belakang diatas, maka dapat diambil beberapa rumusan masalah, yaitu:

a. Bagaimana persepsi publik terhadap layanan Starlink berdasarkan hasil klasifikasi sentimen?
b. Bagaimana membangun model prediksi sentimen berdasarkan komentar pengguna yang telah dianalisis?
c. Sejauh mana algoritma Random Forest dapat mengotomatisasi proses klasifikasi dan prediksi sentimen pada komentar terkait dengan Starlink

1.3 Batasan Masalah

Adapun batasan masalah yang digunakan penulis dalam penelitian ini adalah:

a. Penelitian tidak mencakup pengambilan keputusan bisnis berdasarkan hasil analisis

sentimen

b. Penelitian tidak membahas strategi monitoring reputasi atau manajemen brand

c. Penelitian tidak mengupas secara mendalam pengalaman pelanggan dalam konteks layanan atau kepuasan pelanggan.

1.4 Tujuan Penelitian

Penelitian ini bertujuan untuk:

a. Menggali informasi opini publik terhadap layanan Starlink di Indonesia melalui analisis sentimen komentar pengguna di platform YouTube.
b. Mengidentifikasi persepsi publik terhadap layanan Starlink berdasarkan hasil klasifikasi sentimen (positif, negatif, atau netral).
c. Membangun model prediksi sentimen menggunakan algoritma Random Forest berdasarkan data yang telah diklasifikasikan secara otomatis melalui pendekatan *lexicon-based*.
d. Mengevaluasi kemampuan algoritma Random Forest dalam mengklasifikasikan dan memprediksi sentimen dari komentar pengguna secara otomatis dan efisien.

1.5 Manfaat Penelitian

Berdasarkan rumusan masalah, dalam penelitian ini ada beberapa tujuan yang ingin dicapai yaitu:

a. Hasil penelitian ini dapat memberikan gambaran mengenai persepsi public terkait layanan internet *Starlink*, sehingga dapat menjadi bahan pertimbangan sebelum memutuskan untuk membeli dan menggunakan layanan tersebut.

b. Hasil analisis sentimen ini dapat menjadi masukan untuk memahami respons dan kebutuhan pengguna, khususnya yang bersumber dari komentar – komentar public di media social.

c. Penelitian ini menunjukkan bagaimana opini public dari media social dapat diekrstarksi dan diolah menjadi informasi yang berguna untuk mendukung pengambilan keputusan berbasis data.

1.6 Metode Penelitian

Penelitian ini dilaksanakan melalui beberapa metode dari awal hingga akhir untuk mencapai tujuan penelitian yang telah ditentukan. Berikut adalah penjelasan umum mengenai metode yang digunakan dalam penelitian ini:

a. Observasi

   Tahap awal penelitian ini adalah melakukan observasi atau pengamatan terhadap data yang tersedia di *platform* YouTube. Peneliti mengkaji video – video yang berkaitan dengan topik Starlink di Indonesia, khususnya video yang pertama kali membahas topik tersebut dan yang memiliki jumlah *views* terbanyak.

b. Studi Literatur

   Studi literatur dilakukan untuk mendapatkan landasan teori yang kuat terkait dengan analisis sentimen. Fokus utama adalah mengumpulkan informasi dari berbagai sumber terkait, seperti jurnal ilmiah dan publikasi. Secara spesifik, penelitian ini mendalami prinsip – prinsip analisis sentimen serta mencakup pemahaman tentang metodologi dan metrik evaluasi untuk mengukur model klasifikasi.

c. Implementasi Model dan Pengujian Awal

   Tahap ini adalah inti dari penelitian, di mana model analisis sentimen dirancang dan diimplementasikan. Data teks yang telah melalui tahap *preprocessing* dan diubah menjadi representasi numerik akan digunakan untuk melatih model Random Forest. Setelah model dilatih, model diujikan terlebih dahulu menggunakan data uji dan data latih. Proses ini dilakukan untuk memastikan bahwa model dapat mengklasifikasikan sentimen secara otomatis dan akurat sebelum digunakan untuk analisis yang lebih luas.

d. Evaluasi dan Analisis Hasil

   Setelah pengujian awal selesai, tahap selanjutnya adalah evaluasi kinerja model. Hasil pengujian dianalisis secara mendalam menggunakan matriks evaluasi seperti akurasi, presisi, *recall*, dan *F1-score*. Berdasarkan hasil evaluasi ini, model akan digunakan untuk memprediksi sentimen pada kalimat baru. Tahap terakhir adalah menganalisis hasil prediksi tersebut untuk menarik kesimpulan mengenai efektivitas model Random Forest dalam kasus analisis sentimen Starlink.

1.7 Sistematika Penelitian

Sistematika penelitian disusun untuk memberi gambaran secara umum mengenai isi dari laporan penelitian. Berikut ini adalah metodologi penelitian dalam laporan ini, yaitu:

a. BAB I Pendahuluan

   Bab ini memuat gambaran awal tentang penelitian, meliputi latar belakang masalah, rumusan masalah, tujuan penelitian, manfaat penelitian, dan sistematika penulisan.

b. BAB II Landasan Teori

   Bab ini berisi teori – teori dan kerangka pemikiran yang relevan dengan topik penelitian. Landasan teori berfungsi sebagai dasar ilmiah yang memperkuat analisis dan pembahasan.

c. BAB III Metodologi Penelitian

   Bab ini berisi pembahasan mengenai tahapan yang akan digunakan dalam penelitian, mulai dari pengumpulan data hingga tahap visualisasi data.

d. BAB IV Hasil & Pembahasan

   Bab ini memaparkan hasil dari pengujian model menggunakan algoritma pada berbagai skenario pembagian data dan menjelaskan evaluasi model serta hasil pengujian model.

e. BAB V Kesimpulan

   Bab ini memuat ringkasan hasil penelitian yang menjawab rumusan masalah, serta saran yang dapat diberikan berdasarkan temuan penelitian.

BAB II
LANDASAN TEORI DAN KAJIAN PUSTAKA

2.1 Landasan Teori

Penelitian ini didasarkan pada beberapa konsep dan teori yang mendukung analisis serta interpretasi data secara sistematis. Landasan teori bertujuan untuk memberikan kerangka pemikiran yang kuat dalam memahami fenomena yang diteliti serta membantu dalam pemilihan metode yang tepat. Adapun teori-teori yang digunakan dalam penelitian ini sebagai berikut:

2.1.1 Analisis Sentimen

Analisis sentimen, yang juga dikenal sebagai *opinion mining*, adalah pendekatan berbasis komputasi yang digunakan untuk mengevaluasi opini, persepsi, dan emosi seseorang terhadap suatu objek tertentu. Objek atau entitas yang dimaksud dapat berupa tokoh, kejadian, maupun isu tertentu, yang umumnya dianalisis melalui data ulasan atau tanggapan (Medhat et al., 2014). Analisis sentimen merupakan bagian dari sebuah studi komputasi dalam *Natural Language Processing* (NLP) yang bertujuan untuk mengekstrak nilai sentimen dari teks. Sumber teks ini dapat berasal dari dokumen, ulasan produk, berita, media sosial, dan yang lainnya. Nilai sentimen dapat berupa pendapat atau emosi terhadap sebuah entitas. Nilai tersebut bisa dikategorikan sebagai negatif, netral, atau positif. (Ahmad & Gata, 2022).

2.1.2 Klasifikasi Teks

Klasifikasi teks merupakan proses pengelompokan teks atau dokumen ke dalam berbagai kategori atau kelas tertentu. Secara umum, proses ini melibatkan beberapa tahap, seperti ekstraksi fitur, reduksi dimensi, pemilihan algoritma pengklasifikasi, serta evaluasi performa model (Widhiyasana et al., 2021). Klasifikasi teks (*text classification)* telah menjadi subjek penelitian yang luas di bidang *Natural Processing Language* (NLP). Penerapan klasifikasi teks telah banyak digunakan dalam kegiatan sehari – hari (Ajijah & Kurniawan, 2023).

2.1.3 Natural Language Processing (NLP)

Natural Language Processing (NLP) mengacu pada metode komputasi yang digunakan untuk menganalisis data teks. Untuk mencapai kemampuan pemrosesan bahasa yang menyerupai manusia dalam berbagai aktivitas dan aplikasi, *NLP* mencakup kumpulan teknik berbasis komputasi yang memiliki landasan teoretis. Teknik – Teknik tersebut diterapkan untuk mengevaluasi dan merepresentasikan teks yang muncul secara alami, baik pada tingkat analisis linguistik tunggal maupun tingkat analisis linguistik majemuk. Bidang *NLP* mencakup

berbagai tantangan penelitian dan metodologi. Tugas – tugas *NLP* meliputi berbagai aplikasi, seperti peringkasan teks digital, mesin penjawab pertanyaan, pencarian informasi, dan penerjemahan (Epoka, 2023). NLP menawarkan sejumlah keunggulan, seperti kemampuannya dalam mengolah data teks dalam jumlah besar serta mengotomatiskan berbagai tugas yang berkaitan dengan bahasa, sehingga dapat meningkatkan efisiensi dan produktivitas (Paula et al., 2024).

2.1.4 Text Mining

Text mining merupakan proses ekstraksi pola yang bertujuan untuk mendapatkan informasi dan pengetahuan dari berbagai sumber data yang tidak terstruktur. Meskipun *text mining* dan *data mining* memiliki tujuan yang sama, keduanya menggunakan proses yang serupa tetapi berbeda dalam jenis inputnya. *Input* untuk *text mining* mencakup data tidak terstruktur seperti dokumen, kata, file *PDF*, kutipan, komentar, dan ulasan, sedangkan *data mining* menggunakan data terstruktur (Muhammad, 2019). Dalam konteks penelitian, data teks tidak terstruktur dalam jumlah besar dapat diperoleh dari berbagai sumber seperti literatur ilmiah, transkrip pada sektor pemasaran dan ekonomi, pidato politik seperti kampanye presiden dan pidato pelantikan, serta notulen rapat. Selain itu, sumber daring seperti email, halaman web, blog, media sosial, dan komentar pengguna juga menjadi sumber informasi yang kaya dan sering dimanfaatkan dalam proses text mining (Hassani et al., 2020).

2.1.5 Web Scraping

Web scraping adalah teknik yang digunakan untuk secara otomatis mengambil informasi dari sebuah situs *web* tanpa perlu menyalinnya secara manual. Tujuan dari *web scraping* adalah mencari jenis informasi tertentu, mengekstraknya, dan menggabungkannya ke dalam halaman web baru (Vargiu & Urru, 2012).

2.1.6 Random Forest

Random Forest adalah suatu metode *ensemble* yang terdiri dari sejumlah *decision tree* acak, yang menghasilkan prediksi dengan menggabungkan hasil prediksi dari masing – masing pohon keputusan tersebut (Vens, 2013). Metode ini menggunakan pohon keputusan *(decision tree)* secara pararel, yang diperkenalkan oleh T. Kam Ho pada tahun 1995 (Kowsari et al., 2019). Setiap pohon keputusan dilatih menggunakan sampel data yang berbeda, dan pemilihan atribut dilakukan secara acak pada setiap pohon yang dibentuk.

Metode *random forest* memiliki beberapa keunggulan, antara lain kemampuan meningkatkan akurasi meskipun terdapat data yang hilang, ketahanan terhadap data pencilan *(outlier)*, serta efisiensi dalam hal penyimpanan data. Selain itu, *random forest* juga dilengkapi

dengan proses seleksi fitur yang mampu memilih atribut terbaik, sehingga dapat meningkatkan kinerja model klasifikasi (Manullang et al., n.d.).

2.1.7 Lexicon-Based Approach

Pendekatan berbasis leksikon didasarkan pada anggapan bahwa makna suatu teks dapat ditentukan dari polaritas kata atau frasa yang terkandung di dalamnya. Dengan kata lain, teks dianggap memiliki sentimen tertentu karena terdiri atas kata – kata yang membawa muatan positif, maupun negatif (Catelli et al., 2022).

2.2 Kajian Pustaka

Dalam penulisan skripsi ini peneliti mencari informasi dari beberapa penelitian dan jurnal sebelumnya sebagai acuan poin penting untuk perbandingan, seperti algoritma, dataset, dan hasil yang diberikan pada Tabel 2.1.

Tabel 2.1 Kajian Pustaka

Judul	Publikasi	*Dataset*	Algoritma	Hasil
ANALISIS SENTIMEN APLIKASI BRIMO PADA ULASAN PENGGUNADI GOOGLE PLAY MENGGUNAKAN ALGORITMA NAIVE BAYES	(Khoirul et al., 2023)	*Google Play Store* - Ulasan	*Naïve Bayes Classification*	Dari 1550 data teks, didapatkan hasil sebanyak 1012 atau 53,09% sentimen positf dan sentimen negatif sebanyak 894 atau 46,90%. Pada penelitian ini, mendapatkan hasil *accuracy* sebesar 84,52%, *precision* 82,51%, dan *recall* 87,62%.
Perbandingan Metode K-NN Dan Metode Random Forest Untuk Analisis Sentimen pada Tweet Isu Minyak Goreng di Indonesia	(Yanti et al., 2023)	*Twitter - Tweet*	*K-Nearest Neighbor (KNN)* dan *Random Forest*	Penelitian tersebut membandingkan dua algoritma berbeda lalu membaginya dengan tiga skenario. Skenario pertama berjumlah 500 data, skenario kedua berujumlah 800 data, dan skenario ketiga berjumlah 1.000 data dimana menggunakan rasio data latih dan data uji 80:20. Hasilnya adalah hasil pengujian paling rendah yaitu metode K-NN pada skenario pertama dengan akurasi sebesar 71,50%, presisi 47,83%, dan recall 37,45%. Sementara hasil pengujian paling tinggi yaitu metode K-NN pada skenario kedua dengan akurasi 74,58%, presisi 56,75%, dan recall 44,57%. Sedangkan, untuk rata – rata hasil pengujian dari metode K-NN yaitu akurasi 72,86%, presisi 52,26% dan recall 41,04%. Sedangkan rata – rata hasil pada metode Random Forest yaitu akurasi 73,37%, presisi 52,26% dan recall 34,28%.

SENTIMEN ANALISIS PERILAKU PENGGEMAR COLDPLAY DI MEDIA SOSIAL TWITTER MENGGUNAKAN METODE NAIVE BAYES	(Fudholi et al., 2024)	*Twitter - Tweet*	*Naïve Bayes Classification*	Penelitian tersebut membuktikan bahwa model mampu memprediksi dengan tingkat akurasi sebesar 80,25%. Tingkat akurasi ini bergantung pada data yang diperoleh dari unggahan para penggemar *Coldplay* di *Twitter*. Model berhasil mendeteksi 29 tindakan yang bersifat positif, dan 36 yang bersifat negatif. Namun, terdapat kekeliruan dalam klasifikasi, yang di mana 4 tindakan negatif teridentifikasi sebagai positif, dan 12 tindakan positif teridentifikasi sebagai negatif. Hal ini menunjukkan bahwa adanya bias dalam prediksi.
Sentiment Analysis in Indonesian Healthcare Applications using IndoBERT Approach	(Imaduddin et al., n.d.)	*Google Play Store* - Ulasan	*IndoBERT*	Dalam penelitian tersebut sebanyak 9.310 data ulasan pengguna dan masing- masing diberi label positif atau negatif. Dalam proses pelatihan ini, digunakan 10 epoch bersamaan dengan Adam's Optimizer serta menggunakan *learning rate* sebesar 1e-6. Proses tersebut menghasilkan nilai akurasi sebesar 96%, F1-Score sebesar 95%, *recall* sebesar 96%, dan *precision* sebesar 95%. Dalam penelitian tersebut juga disebutkan bahwa IndoBERT merupakan pendekatan yang kuat untuk analisis sentiment dengan teks bahasa Indonesia.
Analisis Sentimen Untuk Memprediksi Hasil Calon Pemilu Presiden Menggunakan Lexicon Based dan Random Forest	(Manullang et al., n.d.)	*Twitter - Tweet*	*Random Forest dengan Lexicon Dictionary*	Penelitian ini menggunakan algoritma *Random Forest* dengan pendekatan *lexicon based.* Hasilnya menunjukkann bahwa model tersebut dapat memprediksi sentimen positif sebesar 96%, sentimen netral sebesar 97%, dan sentimen negative sebesar 48%. Secara keseluruhan, model yang digunakan memiliki akurasi sebesar 88%.
TWITTER SENTIMENT ANALYSIS TOWARDS QATAR AS HOST OF THE 2022 WORLD CUP USING TEXTBLOB	(S. Dewi & Arianto, n.d.)	*Twitter - Tweet*	*TextBlob*	Pada penelitian ini, data sebanyak 244.832 tweets didapatkan melalui web scrapping dengan cara pencarian kata kunci "2022 FIFA World Cup in Qatar". Penelitian ini menggunakan tiga tahapan, tahap pertama sebelum Qatar terpilih menjadi tuan rumah terdapat 88,46% sentiment positif dan 11,54% sentiment negatif. Tahap kedua yaitu setelah Qatar terpilih menjadi tuan rumah terdapat 79,38% sentiment positif dan 20,62% sentiment negatif. Dan

				yang terakhir tahap ketiga ketika Piala Dunia 2022 berlangsung di Qatar terdapat 83,72% sentiment positif dan 16,38% sentiment negatif. Berdasarkan penelitian tersebut, diperoleh hasil akurasi sebesar 83%.
Sentimen Analisis Pandangan Masyarakat Terhadap Vaksinasi Covid 19 Menggunakan K-Nearest Neighbors	(Apriliani et al., 2023)	*Twitter - Tweet*	*K-Nearest Neighbors (KNN)*	Pada penelitian ini, data diambil dengan cara *crawling* data di sosial media bernama *Twitter* dengan menggunakan kata kunci "Vaksin di *Booster*" dan memperoleh sebanyak 2.241. Pada penelitian ini, penulis juga membandingkan hasil analisis sentimen menggunakan algoritma lain seperti *Naïve Bayes Classification*, dan *Support Vector Machine*, namun hasil tertinggi ada pada algoritma *K-Nearest Neighbors (KNN)* dengan parameter k = 5 dan K-Fold 10 dan menghasilkan nilai akurasi sebesar 79,25%
ANALISIS SENTIMEN APLIKASI SHOPEE, TOKOPEDIA, LAZADA DAN BLIBLI MENGGUNAKAN LEKSIKON DAN RANDOM FOREST	(Syah et al., 2024)	*Google Play Store* – Ulasan	*Random Forest* dengan *Lexicon Dictionary*	Pada penelitian ini, sentimen analisis dilakukan menggunakan algoritma Random Forest dengan pendekatan Lexicon-based. Penelitian ini menghasilkan bahwa aplikasi Lazada memiliki kinerja terbaik dengan akurasi 88,33%, presisi 88,88%, recall 88,33%, dan F1 score 88,34%. Aplikasi Blibli berada di posisi kedua dengan akurasi 85,66%, presisi 85,82%, recall 85,66%, dan F1 score 85,60%. Shopee memiliki akurasi 85,16%, presisi 85,62%, recall 85,16%, dan F1 score 85,26%. Tokopedia menunjukkan performa terendah dengan akurasi 80,33%, presisi 80,96%, recall 80,33%, dan F1 score 80,12%.
SENTIMEN ANALISIS TERHADAP PUAN MAHARANI SEBAGAI KANDIDAT CALON PRESIDEN 2024 BERDASARKAN OPINI TWITTER MENGGUNAKAN METODE NAIVE BAYES DAN ADABOOST	(A. R. Dewi et al., 2023)	*Twitter - Tweet*	*Naive Bayes Classification* dan *AdaBoost*	Penelitian ini menghasilkan nilai akurasi sebesar 70,5% untuk algoritma *Naïve Bayes Classification*. Sedangkan untuk algoritma *AdaBoost* sebesar 68,4%
Analisis Sentimen Popularitas Aplikasi Moodle Dan Edmodo Menggunakan	(Yolanda et al., 2022)	*Google Play Store* – Ulasan	*Support Vector Machine (SVM)*	Dalam penelitian tersebut, aplikasi *Edmodo* memiliki presentase yang lebih tinggi dibandingkan dengan aplikasi *Moodle*. Hal ini ditunjukkan

Algoritma Support Vector Machine				dengan hasil penelitian tersebut bahwa aplikasi *Edmodo* memiliki nilai akurasi sebesar 84%, nilai *precision* sebesar 93%, *recall* sebesar 82%, dan *f1-score* sebesar 87%. Sedangkan untuk aplikasi *Moodle* memiliki nilai akurasi sebesar 82%, *precision* sebssar 79%, *recall* 100%, dan *f1-score* sebesar 88%.
SENTIMEN ANALISIS TERHADAP CUSTOMER REVIEW PRODUK SHOPEE BERBASIS WORDCLOUD DENGAN ALGORITMA NAÏVE BAYES CLASSIFIER	(Hanafiah et al., 2023)	*Shopee* – Ulasan produk	*Naïve Bayes Classification*	Pada penelitian ini dilakukan 4 kali pengujian menggunakan rasio pembagian data training:data latih yang berbeda, yaitu 60:40, 70:30, 80:20, dan 90:10. Nilai akurasi tertinggi yaitu pada rasio 60:40 yaitu sebesar 0.85 atau 85%.
Analisis Sentimen Relokasi Ibukota Nusantara Menggunakan Algoritma Naïve Bayes dan KNN	(Syahril Dwi Prasetyo et al., 2023)	*Twitter - tweet*	*Naïve Bayes Classification* dan *K-Nearest Neighbor*	Penelitian ini memiliki hasil proses analisis menggunakan algoritma *K-Nearest Neighbor* memiliki nilai akurasi sebesar 88,12% dan diikuti dengan nilai presisi sebesar 93,98% dan *recall* sebesar 81,53%. Sedangkan untuk algoritma *Naïve Bayes Classification* memiliki nilai akurasi sebesar 82,27% diikuti dengan nilai presisi sebesar 86,36% dan nilai *recall* sebesar 76,93%. Ini menandakan bahwa algoritma *K-Nearest Nieghbor* memiliki nilai yang lebih baik dibandingkan dengan algoritma *Naïve Bayes Classification.*
Sentimen Analisis Review Aplikasi Digital Korlantas Pada Google Play Store Menggunakan Metode SVM	(Kaburuan & Setiawan, 2023)	*Google Play Store* – Ulasan	*Support Vector Machine (SVM)*	Pada rasio data 90:10, algoritma *SVM* mendapatkan nilai akurasi tertinggi yaitu, 82%, sedangkan untuk akurasi terendahnya pada rasio 80:20 dan 60:40 keduanya mendapatkan hasil nilai akurasi sebesar 74%

Berdasarkan Tabel 2.1, terdapat kesimpulan lain mengenai penggunaan Random Forest dalam analisis sentimen. Selain keunggulan akurasi, Random Forest juga menonjol karena kemampuannya mengatasi tantangan data yang sering ditemukan dalam analisis sentimen.

Dibandingkan dengan Naïve Bayes, yang dalam satu penelitian menunjukkan adanya bias dalam prediksi (Fudholi et al., 2024), Random Forest memiliki ketahanan yang lebih baik terhadap data yang noise atau tidak terstruktur dengan baik. Algoritma ini membangun banyak pohon keputusan dan menggabungkan hasilnya, sehingga mengurangi risiko overfitting yang bisa terjadi pada satu pohon keputusan.

Selain itu, dalam perbandingan dengan K-NN, Random Forest menunjukkan performa yang relatif lebih stabil. Meskipun akurasinya tidak jauh berbeda, Random Forest tidak terlalu sensitif terhadap data yang tidak seimbang, menjadikannya pilihan yang lebih solid untuk dataset yang bervariasi.

Secara keseluruhan, penggunaan *Random Forest* dalam analisis sentimen lebih dari sekadar menghasilkan akurasi yang tinggi. Algoritma ini dipilih karena stabilitasnya, kemampuannya mengurangi bias, dan ketahanannya terhadap kompleksitas data yang sering ada dalam ulasan atau unggahan di media sosial.

BAB III
METODOLOGI PENELITIAN

Penelitian ini dirancang untuk menganalisis sentimen dan membuat model prediksi sentimen dengan mengikuti lima tahapan utama. Tahap pertama adalah collection data, di mana data akan dikumpulkan dari sumber yang telah ditentukan. Data tersebut berupa teks komentar netizen dari media sosial bernama YouTube yang mengandung ekspresi sentimen.

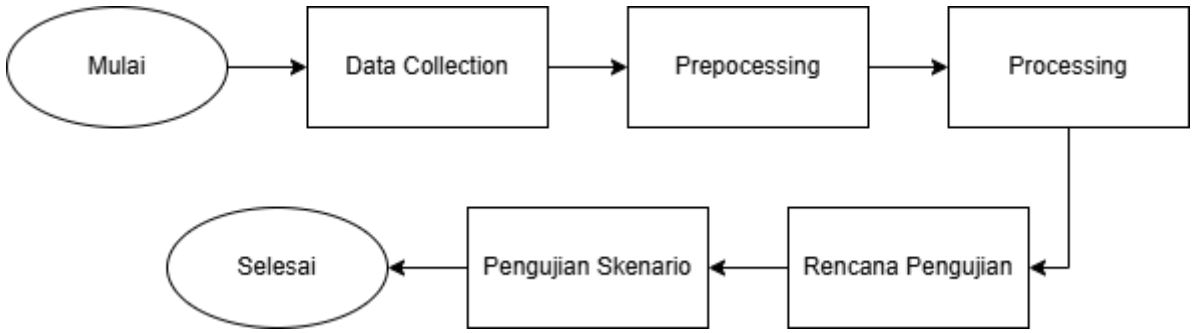
Tahap kedua adalah *preprocessing data*, di mana data yang telah dikumpulkan akan dibersihkan dan diubah menjadi format yang sesuai. Proses ini dimulai dari penghapusan duplikasi data, dan diakhiri dengan *stemming*.

Tahap selanjutnya adalah *processing data*, di mana data akan diproses agar siap untuk dianalisis. Pada tahap ini, nilai sentimen akan diberlakukan pada setiap data, yang nantinya akan digunakan sebagai dasar untuk klasifikasi. Setelah setiap data memiliki nilai sentimen, data akan divisualisasikan menggunakan *wordcloud* per kategori sentimen.

Setelah data mendapatkan nilai sentimennya, selanjutnya masuk ke rencana pengujian. Rencana pengujian ini meliputi pembagian model dengan rasio data latih dan data uji yang berbeda, pembuatan model prediksi sentimen, dan pengujian model prediksi sentimen dengan kalimat baru.

Tahapan terakhir adalah pengujian skenario. Tahapan ini berisi tentang hasil dari tahapan rencana pengujian. Pengujian model dengan berbagai rasio pembagian data latih dan data uji, yaitu 60:40, 70:30, dan 80:20. Pengujian ini bertujuan untuk mengetahui pengaruh jumlah data latih terhadap performa model. Pada setiap rasio, model akan dilatih menggunakan algoritma *Random Forest*. Kinerja model kemudian dievaluasi secara menyeluruh dengan menggunakan *confusion matrix*, yang mencakup metrik akurasi, presisi, recall, dan *F1-score*. Hasil dari pengujian ini akan dibandingkan untuk menemukan rasio pembagian data yang memberikan performa terbaik. Setelah rasio terbaik ditemukan, tahap selanjutnya adalah membangun model prediksi sentimen sederhana menggunakan rasio tersebut. Model ini kemudian akan diuji kembali dengan kalimat – kalimat baru yang tidak ada dalam data pelatihan. Kalimat – kalimat tersebut dipilih secara spesifik untuk merepresentasikan berbagai variasi gaya bahasa, seperti bahasa gaul, kalimat panjang, dan kalimat yang mengandung sarkasme. Pengujian ini sangat penting untuk menilai kemampuan model dalam memahami konteks dan memprediksi sentimen pada data yang bervariasi di dunia nyata. Skenario penelitian ini divisualisasikan

menggunakan diagram alur pada gambar 3.1 yang mencakup dari tahap pengumpulan data hingga pengujian skenario.



Gambar 3.1 Diagram Alur Penelitian

3.1 Data Collection (Pengumpulan Data)

Tahap pengumpulan data merupakan proses mengumpulkan data teks atau ulasan yang akan dianalisis untuk menentukan sentimen (positif, negatif, atau netral). Data yang dikumpulkan biasanya berasal dari sumber – sumber seperti media sosial (misalnya Twitter, Instagram, atau Facebook), platform e-commerce (seperti Tokopedia atau Shopee), forum online, atau situs ulasan (contohnya TripAdvisor atau Google Reviews).

Pada penelitian ini, data yang digunakan adalah komentar *netizen* dari dua video YouTube yang mengulas tentang Starlink di Indonesia, yaitu video yang berjudul "Review Starlink Indonesia – Ini Revolusioner" dari kanal GadgetIn sebagai video tentang Starlink dengan jumlah *views* paling banyak di Indonesia, dan "STARLINK INDONESIA REVIEW, INTERNET DARI LANGIT TANPA SINYAL DIMANAPUN ANDA BERADA (TERLENGKAP)" dari kanal NanangMrk sebagai video yang pertama kali membahas tentang Starlink di Indonesia. Pervideo akan diambil dua kategori, yaitu kategori *top comments* dan kategori *newest* dengan sebanyak maksimal 2.500 data komentar perkategori dikarenakan akses terbatas dalam pengambilan data. Selanjutnya, dataset ini akan diberi nama "Top Comment – GadgetIn", "Newest – GadgetIn", "Top Comment – NanangMrk", dan "Newest – NanangMrk".

Berikut adalah dataset untuk video "Review Starlink Indonesia – Ini Revolusioner" dari kanal GadgetIn seperti pada Gambar 3.2 dan Gambar 3.3.

GadgetIn - Top Comments

[11]:

	comment
0	Starbucks keliling
1	Starlink ini adalah investasi yang sangat bagu...
2	Malu gak sih yang katanya Elite Global dengan ...
3	Muncak bawa starlink buka jasa wifi di sana
4	Bang kenapa manusia saling menyakiti?
...	...
2495	Bang kalo hujan terus ada petir itu aman gak ?...
2496	Cocok banget untuk orang yang punya pulau prib...
2497	untuk di ruko apakah jika di taruh rooftop bs ...
2498	kalo lagi hujan lebat, rawan kena petir gak ba...
2499	boleh minta link tempat pembelian powerbank ya...

2500 rows × 1 columns

Gambar 3.2 Dataset Top Comments dari Kanal GadgetIn

GadgetIn - Newest Comments

[10]:

	comment
0	Kata nya harga starlink hanya 1,999000 jt sj t...
1	Ada bahayanya juga sih..Elon mars mau ledakkan...
2	Bisa dipakai sampe brpa device bang
3	ngeri
4	Bang kenapa manusia saling menyakiti?
...	...
2495	Lagi nabung beli alatnya
2496	Fibers CBN keren, saya pakai itu, tembus 50 Mb...
2497	Fibers CBN keren, saya pakai itu, tembus 50 Mb...
2498	Kumpulin 5 rumah samping rumah untuk patungan....
2499	Kalau di kapal bang

2500 rows × 1 columns

Gambar 3.3 Dataset Newest Comments dari Kanal GadgetIn

Berikut adalah dataset untuk video “STARLINK INDONESIA REVIEW, INTERNET DARI LANGIT TANPA SINYAL DIMANAPUN ANDA BERADA (TERLENGKAP)” dari kanal NanangMrk seperti pada Gambr 3.4 dan Gambar 3.5.

NanangMrk - Top Comments

[13]:

	comment
0	#INDONESIA DARURAT MEMBACA!!!!! \n\nSTARLINK I...
1	Kalau ada yg nanyain Starlink di channelku, ta...
2	Yg bilang mahal. Padahal biaya pembuatan athen...
3	Baru pasang kemaren, emang se worth it itu. ha...
4	Sebenernya kecepatannya Biasa aja kalo di nega...
...	...
1868	Minta no wa nya bang
1869	Mahal
1870	SATELIT APANYA ?🤩 WONG ROKET AJA GA BISA NEMBU...
1871	Paket internet di kota brapa harganya kecepata...
1872	bukan untuk perkotaan, tapi untuk daerah terlu...

1873 rows × 1 columns

Gambar 3.4 Dataset Top Comments dari Kanal NanangMrk

NanangMrk - Newest Comments

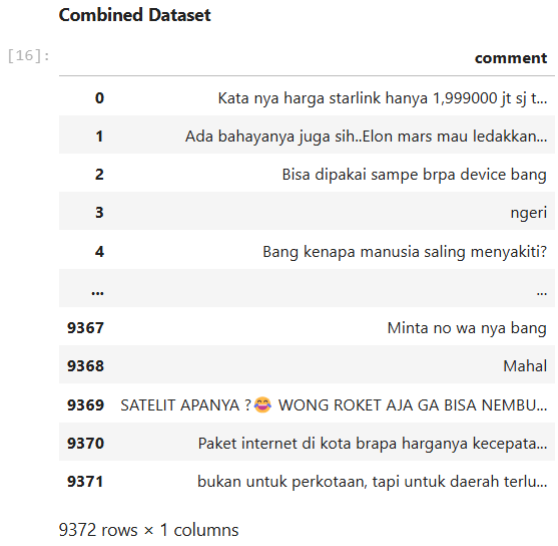
[12]:

	comment
0	#INDONESIA DARURAT MEMBACA!!!!! \n\nSTARLINK I...
1	Berpa hrga Starlink yg bisa di pake buat bisni...
2	Kira kira berapa unit HP yang bisa konek ke di...
3	Jarak maksimal berapa jauh mas signalnya
4	Hrga brpa 1 unitnya starlink yg lg d demo tu ?
...	...
2494	Itu lah akal nya orang kapir, dalam teknologi ...
2495	Bandingin sama sinyal 5G bg
2496	Bang ini up to berapa device yang terhubung?
2497	perbulannya berpah itu.....bisa buat rt rw ne...
2498	Dpet ip publik statis ga ya??

2499 rows × 1 columns

Gambar 3.5 Dataset Newest Comments dari Kanal NanangMrk

Keempat dataset tersebut kemudian digabungkan menjadi satu menjadi sebuah dataset gabungan baru. Setelah digabungkan, dataset baru memiliki total 9.372 data komentar seperti pada Gambar 3.6.



Combined Dataset

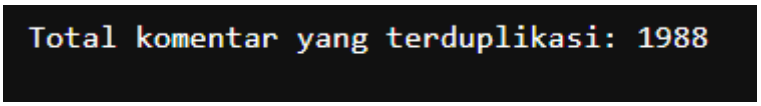
[16]:

	comment
0	Kata nya harga starlink hanya 1,999000 jt sj t...
1	Ada bahayanya juga sih..Elon mars mau ledakkan...
2	Bisa dipakai sampe brpa device bang
3	ngeri
4	Bang kenapa manusia saling menyakiti?
...	...
9367	Minta no wa nya bang
9368	Mahal
9369	SATELIT APANYA ?🤩 WONG ROKET AJA GA BISA NEMBU...
9370	Paket internet di kota brapa harganya kecepata...
9371	bukan untuk perkotaan, tapi untuk daerah terlu...

9372 rows × 1 columns

Gambar 3.6 Dataset Gabungan dari Kedua Video

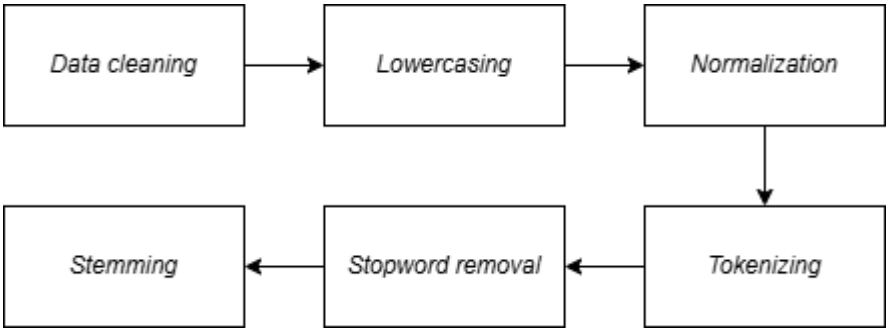
Namun, dalam proses penggabungan dataset ini, terdapat adanya satu atau lebih data komentar yang muncul di dalam dataset *top comments* yang juga ditemukan di dalam dataset *newest comments*. Hal ini menyebabkan terjadinya duplikasi data, dimana komentar yang sama muncul lebih dari sekali dalam dataset gabungan. Setelah dilakukan pengecekan, ditemukan sebanyak 1.988 data yang terduplikasi seperti pada Gambar 3.7. Oleh karena itu, data perlu dibersihkan di tahap selanjutnya yaitu tahap *preprocessing* data.



Gambar 3. 7 Total Komentar Terduplikasi

3.2 Preprocessing Data

Preprocessing data merupakan tahapan awal dalam pengolahan data yang bertujuan untuk menyiapkan data mentah agar dapat digunakan dalam pemodelan. *Preprocessing* biasanya melibatkan pembersihan data dan transformasi data dasar untuk memastikan data sesuai dengan format atau persyaratan *input* dari algoritma yang digunakan dan tahapan *preprocessing* data dapat dilihat pada Gambar 3.8.



Gambar 3.8 Tahapan Prepocessing Data

Tahapan dalam *preprocessing* data pada penelitian meliputi sebagai berikut:

a. *Data Cleaning*

Data *cleaning* merupakan langkah yang penting untuk memastikan bahwa data yang akan diproses berkualitas dan siap digunakan untuk analisis lebih lanjut. Tahapan ini penting untuk meningkatkan kualitas data sebelum data tersebut digunakan dalam analisis atau pemodelan. Data sering kali mengandung kesalahan seperti nilai yang hilang, data duplikat, kesalahan pengetikan, atau inkonsistensi format, yang dapat memengaruhi hasil analisis.

Dalam tahap *data cleaning*, ada dua langkah – langkah, yaitu *delete duplicates* dan *text cleaning*. Berikut adalah masing – masing penjelasannya:

1. *Delete Duplicates*

Menghapus duplikasi data adalah proses mengidentifikasi dan menghilangkan data yang berulang atau identik dalam sebuah kumpulan data. Duplikasi terjadi ketika data yang sama disimpan lebih dari sekali, baik karena kesalahan *input*, penggabungan beberapa sumber data, atau proses pengolahan yang tidak terkontrol. Keberadaan data duplikat dapat berdampak pada kualitas data. Setelah dilakukan penghapusan duplikasi, kini tersisa 7.384 data komentar yang sebelumnya berjumlah 9.372 data komentar seperti pada Gambar 3.9.

[157]:

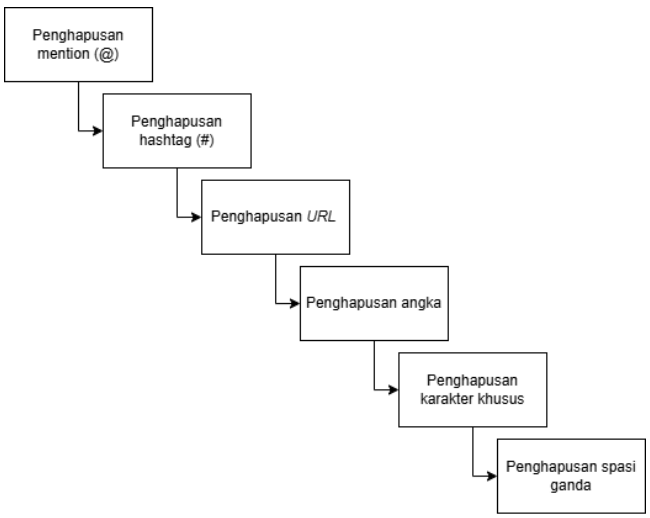
	comment
0	Kata nya harga starlink hanya 1,999000 jt sj t...
1	Ada bahayanya juga sih..Elon mars mau ledakkan...
2	Bisa dipakai sampe brpa device bang
3	ngeri
4	Bang kenapa manusia saling menyakiti?
...	...
8334	@dodollucknutzz loh kan emang sudah dibahas....
8335	@dikky3349 Kan sudah dipindah dari atas mobi...
8336	@dodollucknutzz haha klo cuman dibawa ke ata...
8337	@dodollucknutzz maksudnya, mau ngetes apakah...
8338	Alhamdulillah trimakasih

7384 rows × 1 columns

Gambar 3.9 *Delete Duplicates*

2. *Text Cleaning*

Text cleaning adalah proses membersihkan data teks dari elemen – elemen yang tidak digunakan dalam analisis teks. Tujuannya adalah untuk memastikan bahwa data teks yang akan digunakan dalam analisis atau pemodelan hanya berisi informasi penting. Proses *text cleaning* dapat divisualisasikan dalam bentuk diagram seperti pada Gambar 3.10.



Gambar 3.10 *Diagram Alur Text Cleaning*

Terdapat enam langkah – langkah dari tahapan *text cleaning.* Berikut adalah penjelasan langkah – langkah dari *text cleaning* yaitu:

- Penghapusan *Mention (@mention)*

  Tanda *mention* yang kemudian diikuti oleh nama pengguna/*author* di dalam sebuah sosial media, yang secara otomatis menandai atau menyebut pengguna/*author* tertentu dalam sebuah percakapan. Dalam konteks analisis teks, seperti analisis sentiment maupun klasifikasi, *mention* tidak memiliki bobot yang signifikan. Dikarenakan *mention* lebih berfungsi sebagai *metadata*, bukan sebagai bagian dari isi utama teks yang mencerminkan opini, emosi, ataupun ekspresi dari pengguna. Oleh karena itu, *mention* dihapus didalam proses *text cleaning* agar mengurangi *noise*.

- Penghapusan *Hashtag (#hashtag)*

  Hashtag adalah fitur yang juga sering muncul dalam komentar pada sosial media, baik itu di Instagram, Twitter, Facebook, YouTube, dan lain sebagainya. *Hashtag* ini ditandai dengan simbol "#" dan diikuti oleh kata yang menjadi topik utama, seperti "#starlink" contohnya, sehingga *hashtag* biasanya digunakan untuk memperkuat topik pembahasan yang sedang dibahas. Namun, sering kali *hashtag* hanya berfungsi sebagai *metadata*, yang bertarti tidak memberikan informasi tambahan dan tidak memiliki bobot yang kuat. Oleh karena itu, penghapusan *hashtag* diperlukan untuk membersihkan data dari elemen – elemen yang tidak langsung berhubungan dengan klasifikasi sentimen.

- Penghapusan *URL (Uniform Resource Locator)*

  Komentar di YouTube terkadang menyertakan *URL*, baik itu untuk menautkan ke video lain, artikel, atau konten eksternal yang relevan dengan topik video. Meskipun bagi pengguna tautan *URL* tersebut penting, namun *URL* tidak memberikan informasi langsung mengenai opini yang disampaikan dalam komentar. Maka dari itu, *URL* hanya menjadi sebatas *metadata* yang tidak dibutuhkan dalam proses klasifikasi sentiment, sehingga *URL* dihapus dari data teks untuk menjaga relevansi informasi.

- Penghapusan Angka

  Penghapusan angka dalam klasifikasi sentimen merupakan langkah penting untuk meningkatkan akurasi analisis teks. Dengan menghilangkan

angka, kita dapat fokus pada kata-kata yang lebih relevan dan bermakna dalam konteks emosi yang diekspresikan. Hal ini membantu mengurangi kebisingan data, yaitu informasi yang tidak relevan yang dapat mengganggu proses analisis. Ketika angka dihapus, model klasifikasi dapat lebih efektif dalam menangkap nuansa dan makna sentimen yang terkandung dalam teks, sehingga menghasilkan analisis yang lebih tepat.

Selain itu, proses penghapusan angka juga berkontribusi pada pembersihan data secara keseluruhan. Dengan membersihkan dataset dari elemen – elemen yang tidak diperlukan, dapat memastikan bahwa model klasifikasi bekerja dengan data yang lebih bersih dan terfokus. Ini tidak hanya meningkatkan keakuratan hasil, tetapi juga memudahkan dalam interpretasi hasil analisis. Dengan demikian, penghapusan angka menjadi bagian penting dalam mempersiapkan data untuk klasifikasi sentimen yang lebih efektif.

- Penghapusan Karakter Khusus atau Non-Alfa Numerik

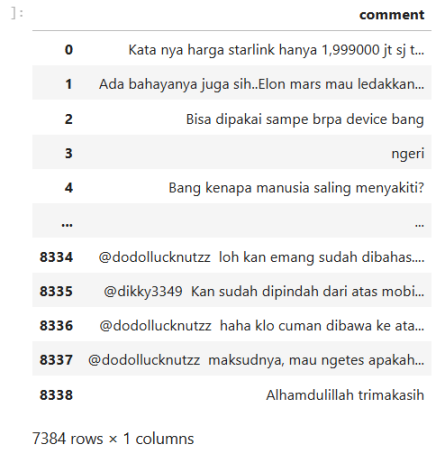
Karakter khusus dan simbol non-alfa numerik seperti angka, tanda baca, atau simbol seperti &, !, %, dan lain – lain sering sekali muncul di dalam komentar sosial media. Terkadang, simbol ini digunakan untuk menegaskan sesuatu atau sebagai cara pengguna penulis (contohnya penggunaan kata “dan” diganti menggunakan simbol “&”). Namun, simbol – simbol tersebut jarang sekali memiliki bobot dalam klasifikasi sentimen. Dengan cara ini, karakter emotikon seperti “😄” dan “💘” juga akan otomatis terhapus karena mereka termasuk dalam kategori non-alfanumerik. Cara ini hanya menghapus karakter emotikon itu sendiri, dan tidak memengaruhi kata – kata yang mendahului atau mengikutinya. Dengan demikian, kata – kata yang relevan tetap dipertahankan, sementara karakter yang tidak memiliki bobot analitis dihilangkan.

- Penghapusan Spasi Ganda

Spasi ganda dapat membuat teks tidak terlihat rapi dan dapat mempengaruhi proses tokenisasi dalam analisis teks. Oleh karena itu, menghapus spasi ganda dilakukan agar menjaga keteraturan teks.

b. *Lowercasing*

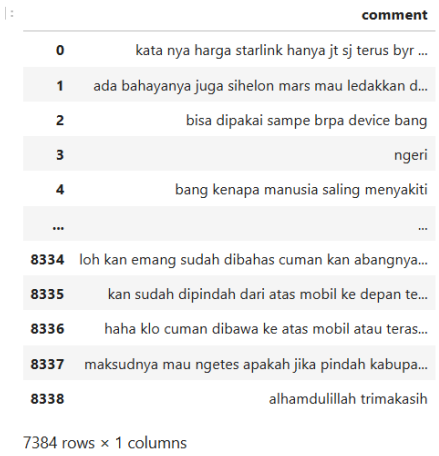
Lowercasing adalah proses mengubah semua huruf kapital dalam teks menjadi huruf kecil atau yang dapat disebut dengan *lowercase.* Dalam *Natural Language Processing (NLP), lowercasing* merupakan salah satu teknik *preprocessing* yang umum digunakan untuk menjaga konsistensi dalam analisis data teks. Selain itu, *lowercasing* dapat meningkatkan kualitas analisis teks karena kata – kata yang sama tidak akan dianggap berbeda hanya karena perbedaan huruf besar/kecil. Gambar 3.11 dan 3.12 menunjukkan dataset sebelum dan sesudahnya dibersihkan serta diberlakukannya *lowercasing.*



	comment
0	Kata nya harga starlink hanya 1,999000 jt sj t...
1	Ada bahayanya juga sih..Elon mars mau ledakkan...
2	Bisa dipakai sampe brpa device bang
3	ngeri
4	Bang kenapa manusia saling menyakiti?
...	...
8334	@dodollucknutzz loh kan emang sudah dibahas....
8335	@dikky3349 Kan sudah dipindah dari atas mobi...
8336	@dodollucknutzz haha klo cuman dibawa ke ata...
8337	@dodollucknutzz maksudnya, mau ngetes apakah...
8338	Alhamdulillah trimakasih

7384 rows × 1 columns

Gambar 3.11 Sebelum dibersihkan dan *lowercasing*



	comment
0	kata nya harga starlink hanya jt sj terus byr ...
1	ada bahayanya juga sihelon mars mau ledakkan d...
2	bisa dipakai sampe brpa device bang
3	ngeri
4	bang kenapa manusia saling menyakiti
...	...
8334	loh kan emang sudah dibahas cuman kan abangnya...
8335	kan sudah dipindah dari atas mobil ke depan te...
8336	haha klo cuman dibawa ke atas mobil atau teras...
8337	maksudnya mau ngetes apakah jika pindah kabupa...
8338	alhamdulillah trimakasih

7384 rows × 1 columns

Gambar 3.12 Sesudah dibersihkan dan *lowercasing*

c. *Normalization*

Normalization atau normalisasi adalah proses awal dalam pemrosesan teks untuk menyamakan teks mentah sehingga lebih konsisten dan mudah dianalisis. Proses ini melibatkan pengubahan teks menjadi bentuk baku. Dalam bahasa Indonesia, normalisasi juga sering mencakup penanganan kata – kata tidak baku atau *slang* agar menjadi kata baku.

Pada sebuah *dataset* tentu saja banyak kata – kata singkatan, *typo*, maupun *slang* atau kata – kata gaul yang dapat mempengaruhi hasil analisis. Maka dari itu, normalisasi sangat penting dalam meningkatkan kualitas analisis. Tanpa normalisasi, bentuk – bentuk kata yang bervariasi ini dapat membuat analisis lebih rumit dan menghasilkan hasil yang tidak akurat. Dengan kata yang sudah

dinormalisasi, algoritma dapat mengenali pola yang konsisten dan melakukan analisis lebih efektif. Gambar 3.13 dan 3.14 menunjukkan dataset sebelum dan sesudah diberlakukan *normalization.*

	comment
0	kata nya harga starlink hanya jt sj terus byr ...
1	ada bahayanya juga sihelon mars mau ledakkan d...
2	bisa dipakai sampe brpa device bang
3	ngeri
4	bang kenapa manusia saling menyakiti
...	...
8334	loh kan emang sudah dibahas cuman kan abangnya...
8335	kan sudah dipindah dari atas mobil ke depan te...
8336	haha klo cuman dibawa ke atas mobil atau teras...
8337	maksudnya mau ngetes apakah jika pindah kabupa...
8338	alhamdulillah trimakasih

7384 rows × 1 columns

Gambar 3.13 Sebelum Normalisasi

	comment
0	kata nya harga starlink hanya juta saja terus ...
1	ada bahayanya juga sihelon mars mau ledakkan d...
2	bisa dipakai sampe berapa device bang
3	ngeri
4	bang kenapa manusia saling menyakiti
...	...
8334	loh kan emang sudah dibahas cuman kan abangnya...
8335	kan sudah dipindah dari atas mobil ke depan te...
8336	haha kalau cuman dibawa ke atas mobil atau ter...
8337	maksudnya mau ngetes apakah jika pindah kabupa...
8338	alhamdulillah trimakasih

7384 rows × 1 columns

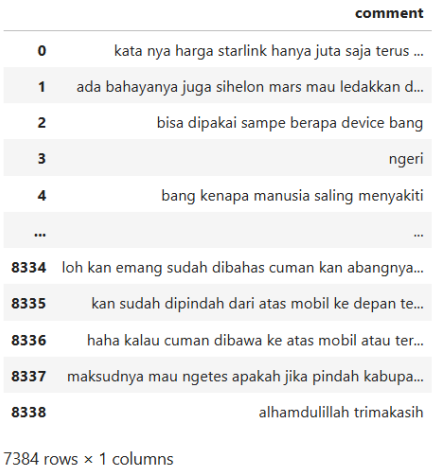
Gambar 3.14 Sesudah Normalisasi

Dataset telah berhasil melalui proses normalisasi. Dapat dilihat sebelum melalui proses normalisasi pada *index* ke 0, kalimat “kata nya harga starlink hanya jt sj terus byr…” berubah menjadi “kata nya harga starlink hanya juta saja terus…”. Setelah tahap normalisasi, *dataset* akan masuk ke tahap selanjutnya, yaitu *tokenizing*,

d. *Tokenization*

Tokenization atau biasa yang disebut dengan tokenisasi adalah proses memecah teks menjadi kata – kata individual yang lebih kecil yang disebut "token." Tokenisasi merupakan langkah penting dalam pemrosesan teks karena dapat membuat algoritma untuk memahami elemen dasar dari teks yang akan dianalisis.

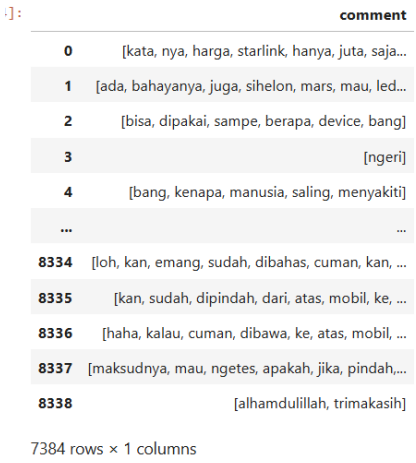
Tokenisasi membantu dalam mempersiapkan teks untuk analisis lebih lanjut, seperti penghitungan frekuensi kata. Setelah melalui proses *tokenizing*, awal kalimat tiap baris pada dataset akan diawali dan diakhiri dengan tanda baca kurung siku ([…]) dan dipecah menjadi kata – kata individual yang dipisahkan menggunakan tanda baca koma (,). Gambar 3.15 dan 3.16 menunjukkan dataset sebelum dan sesudah diberlakukan *tokenizing*.



	comment
0	kata nya harga starlink hanya juta saja terus ...
1	ada bahayanya juga sihelon mars mau ledakkan d...
2	bisa dipakai sampe berapa device bang
3	ngeri
4	bang kenapa manusia saling menyakiti
...	...
8334	loh kan emang sudah dibahas cuman kan abangnya...
8335	kan sudah dipindah dari atas mobil ke depan te...
8336	haha kalau cuman dibawa ke atas mobil atau ter...
8337	maksudnya mau ngetes apakah jika pindah kabupa...
8338	alhamdulillah trimakasih

7384 rows × 1 columns

Gambar 3.15 Sebelum *Tokenizing*



	comment
0	[kata, nya, harga, starlink, hanya, juta, saja...
1	[ada, bahayanya, juga, sihelon, mars, mau, led...
2	[bisa, dipakai, sampe, berapa, device, bang]
3	[ngeri]
4	[bang, kenapa, manusia, saling, menyakiti]
...	...
8334	[loh, kan, emang, sudah, dibahas, cuman, kan, ...
8335	[kan, sudah, dipindah, dari, atas, mobil, ke, ...
8336	[haha, kalau, cuman, dibawa, ke, atas, mobil, ...
8337	[maksudnya, mau, ngetes, apakah, jika, pindah,...
8338	[alhamdulillah, trimakasih]

7384 rows × 1 columns

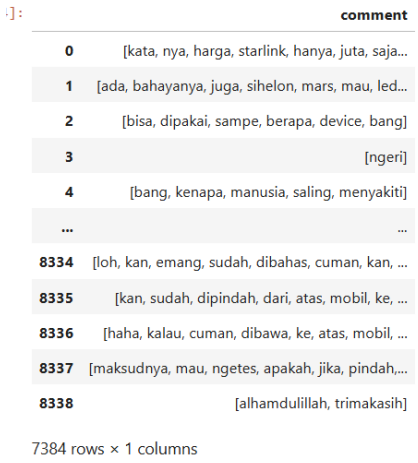
Gambar 3.16 Sesudah *Tokenizing*

e. *Stopword Removal*

Stopword removal adalah proses menghapus kata – kata umum yang sering muncul tetapi biasanya tidak memiliki makna dalam analisis, seperti "dan," "di," "yang," atau "untuk" dalam bahasa Indonesia. Kata – kata ini, yang disebut dengan *stopwords*. *Stopwords* sering kali tidak memberikan informasi penting dalam analisis sentimen atau teks, sehingga penghapusannya membantu mengurangi *noise* dalam data dan mempercepat proses analisis.

Stopword removal berguna dalam meningkatkan akurasi model dengan menghilangkan kata – kata yang tidak digunakan, sehingga model lebih fokus pada kata – kata penting yang benar – benar berguna terhadap makna teks.

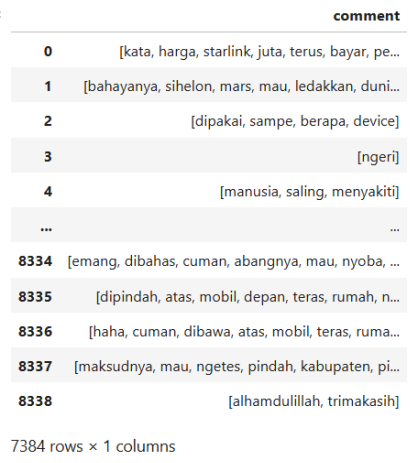
Pada tahap *stopword removal*, penulis menggunakan *library* yang bernama *Sastrawi* untuk pemrosesan teks berbahasa Indonesia. *Library* ini sangat berguna dalam berbagai aplikasi pemrosesan bahasa alami *(NLP)*, seperti analisis sentimen, pencarian informasi, dan klasifikasi teks. Gambar 3.17 dan 3.18 menunjukkan dataset sebelum dan sesudah diberlakukan *stopword removal.*



	comment
0	[kata, nya, harga, starlink, hanya, juta, saja...
1	[ada, bahayanya, juga, sihelon, mars, mau, led...
2	[bisa, dipakai, sampe, berapa, device, bang]
3	[ngeri]
4	[bang, kenapa, manusia, saling, menyakiti]
...	...
8334	[loh, kan, emang, sudah, dibahas, cuman, kan, ...
8335	[kan, sudah, dipindah, dari, atas, mobil, ke, ...
8336	[haha, kalau, cuman, dibawa, ke, atas, mobil, ...
8337	[maksudnya, mau, ngetes, apakah, jika, pindah,...
8338	[alhamdulillah, trimakasih]

7384 rows × 1 columns

Gambar 3.17 Sebelum *Stopword Removal*



	comment
0	[kata, harga, starlink, juta, terus, bayar, pe...
1	[bahayanya, sihelon, mars, mau, ledakkan, duni...
2	[dipakai, sampe, berapa, device]
3	[ngeri]
4	[manusia, saling, menyakiti]
...	...
8334	[emang, dibahas, cuman, abangnya, mau, nyoba, ...
8335	[dipindah, atas, mobil, depan, teras, rumah, n...
8336	[haha, cuman, dibawa, atas, mobil, teras, ruma...
8337	[maksudnya, mau, ngetes, pindah, kabupaten, pi...
8338	[alhamdulillah, trimakasih]

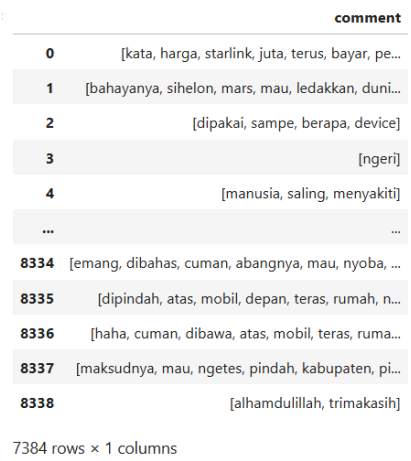
7384 rows × 1 columns

Gambar 3.18 Sesudah *Stopword Removal*

f. *Stemming*

Stemming adalah proses mengubah kata menjadi bentuk dasar, misalnya mengubah kata "berlari" atau "berlarian" menjadi "lari". Dalam analisis teks, *stemming* bertujuan untuk menyederhanakan kata – kata yang beragam bentuknya tetapi memiliki arti yang sama, sehingga meningkatkan konsistensi data.

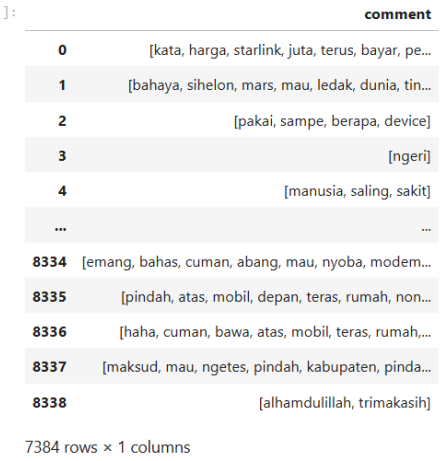
Stemming membantu model untuk tidak terganggu oleh variasi morfologi kata yang sebenarnya tidak memengaruhi makna inti. Dalam analisis sentimen, *stemming* penting karena kata dengan makna yang sama tetapi bentuk berbeda harus dikenali sebagai satu konsep agar analisis lebih akurat. Dengan mengurangi kata ke bentuk dasarnya, stemming juga mengurangi kompleksitas model karena jumlah kata yang dianalisis berkurang, membuat analisis teks lebih efisien dan efektif. Gambar 3.19 dan 3.20 menunjukkan dataset sebelum dan sesudah diberlakukan *stemming.*



	comment
0	[kata, harga, starlink, juta, terus, bayar, pe...
1	[bahayanya, sihelon, mars, mau, ledakkan, duni...
2	[dipakai, sampe, berapa, device]
3	[ngeri]
4	[manusia, saling, menyakiti]
...	...
8334	[emang, dibahas, cuman, abangnya, mau, nyoba, ...
8335	[dipindah, atas, mobil, depan, teras, rumah, n...
8336	[haha, cuman, dibawa, atas, mobil, teras, ruma...
8337	[maksudnya, mau, ngetes, pindah, kabupaten, pi...
8338	[alhamdulillah, trimakasih]

7384 rows × 1 columns

Gambar 3.19 Sebelum *Stemming*



	comment
0	[kata, harga, starlink, juta, terus, bayar, pe...
1	[bahaya, sihelon, mars, mau, ledak, dunia, tin...
2	[pakai, sampe, berapa, device]
3	[ngeri]
4	[manusia, saling, sakit]
...	...
8334	[emang, bahas, cuman, abang, mau, nyoba, modem...
8335	[pindah, atas, mobil, depan, teras, rumah, non...
8336	[haha, cuman, bawa, atas, mobil, teras, rumah,...
8337	[maksud, mau, ngetes, pindah, kabupaten, pinda...
8338	[alhamdulillah, trimakasih]

7384 rows × 1 columns

Gambar 3.20 Sesudah *Stemming*

Kata seperti “bahayanya” pada index ke 1, “dipakai” pada index ke 2, dan “dipindah” pada index ke 8.335 telah berubah menjadi bentuk kata dasarnya.

Setelah semua tahapan *preprocessing* selesai dilakukan, selanjutnya penulis menyimpan *dataset* hasil tahapan *preprocessing.*

3.3 Processing Data

Pada *processing data*, data yang sudah melalui tahapan *preprocessing* akan diolah agar bisa digunakan dalam analisis lebih lanjut. Di dalam *processing* data terdapat dua tahapan yaitu pelabelan sentimen dan visualisasi.

A. Pelabelan Sentimen

Pelabelan sentimen adalah proses penentuan kategori emosi atau opini yang terkandung dalam suatu teks. Ini bertujuan untuk mengidentifikasi dan mengklasifikasikan sikap penulis terhadap topik tertentu berdasarkan kata-kata yang digunakan dalam teks.

Pada penelitian ini, pelabelan sentimen akan dibedakan dalam tiga kategori, yaitu positif, negatif, dan netral. *Lexicon dictionary* bahasa Indonesia digunakan untuk memudahkan proses pelabelan data. *Lexicon dictionary* adalah kumpulan – kumpulan kata atau frasa yang telah diberi nilai sentimen. *Lexicon dictionary* ini

digunakan sebagai refrensi untuk mengidentifikasi sentimen dari teks atau kalimat berdasarkan kata – kata yang ada di dalamnya. Berikut adalah contoh tabel *lexicon dictionary* bahasa Indonesia sentimen positif seperti pada Tabel 3.1.

Tabel 3.1 *Lexicon Dictionary* Bahasa Indonesia Sentimen Positif

Kata	**Nilai**
Bagus	+2
Menarik	+4
Mantap	+5

Dan berikut adalah tabel contoh *lexicon dictionary* bahasa Indonesia sentimen negative seperti pada Tabel 3.2.

Tabel 3.2 *Lexicon Dictionary* Bahasa Indonesia Sentimen Negative

Kata	**Nilai**
Ampas	-3
Jelek	-5
Buruk	-5

Penilaian teks menggunakan ***lexicon dictionary*** dilakukan dengan cara mencocokkan kata – kata dalam teks dengan kata – kata yang ada di *lexicon dictionary*. Untuk menentukan sentimen sebuah teks, sistem akan menjumlahkan nilai dari setiap kata yang ditemukan (Hayaty et al., 2020). Setelah mendapatkan total nilai, kalimat tersebut kemudian diklasifikasikan ke dalam kategori sentimen menggunakan algoritma yang sudah ditetapkan (Musfiroh et al., 2021).

Mengacu pada penelitian (Musfiroh et al., 2021), secara umum dinyatakan sebagai berikut:

1. Jika nilai sentimen > 0, maka positif,
2. Jika nilai sentimen < 0, maka negatif

3. Jika nilai sentimen = 0, maka netral.

Berikut adalah data komentar yang telah diberi label sesuai dengan *lexicon dictionary* bahasa Indonesia. Setiap data komentar telah diberi label sentimen sesuai dengan total skor akhir seperti pada Gambar 3.21.

	comment	sentiment	score
0	kata harga starlink juta terus bayar per bulan...	Positive	2
1	bahaya sihelon mars mau ledak dunia tinggal te...	Negative	-9
2	pakai sampe berapa device	Neutral	0
3	ngeri	Negative	-4
4	manusia saling sakit	Negative	-2
...	...	...	...
7379	emang bahas cuman abang mau nyoba modem bawa p...	Negative	-2
7380	pindah atas mobil depan teras rumah nonton ten...	Positive	1
7381	haha cuman bawa atas mobil teras rumah mah lok...	Negative	-9
7382	maksud mau ngetes pindah kabupaten pindah temp...	Negative	-19
7383	alhamdulillah trimakasih	Positive	9

[7384 rows x 3 columns]

Gambar 3.21 Dataset Setelah Diberi Label Sentimen

Jika data tersebut dipecah menurut nilai sentimennya, maka total data yang bernilai positif sebanyak 2.627, total data yang bernilai negatif sebanyak 3.296, dan 1.461 untuk total data yang bernilai netral seperti pada Tabel 3.3.

Tabel 3.3 Total Data Sentimen

Sentimen	**Total Data**
Positif	2.627
Negatif	3.296
Netral	1.461

B. Visualization

Setelah data komentar memiliki label sentimen, kemudian data akan divisualisasikan. Visualisasi data adalah proses menyajikan data dalam bentuk grafis atau visual untuk mempermudah pemahaman, dan analisis. Dalam penelitian ini, *wordcloud* digunakan dalam visualisasi data. *Wordcloud* adalah

visualisasi data berbentuk kumpulan kata yang disusun secara acak dengan ukuran yang bervariasi. Ukuran setiap kata mencerminkan frekuensi dalam suatu *dataset.* Semakin sering kata muncul, semakin besar ukuran kata tersebut (Iqbal & Yusuf, n.d.). Berikut adalah visualisasi kata yang bernilai positif dalam bentuk *wordcloud* seperti pada Gambar 3.22.



Gambar 3.22 Visualisasi Kata yang Bernilai Positif dalam Bentuk *Wordcloud*

Pada *wordcloud* sentimen positif ini menampilkan kata – kata positif yang sering muncul dalam dataset teks, dengan kata seperti "Starlink", "paket", "rumah", dan "mantap" yang dominan. Kata – kata yang lebih besar mencerminkan frekuensi tinggi, menunjukkan bahwa pengguna sering mengapresiasi aspek tertentu dari layanan Starlink. Selain itu, kata seperti "stabil", "langsung", dan "*wifi*" mengindikasikan pengalaman positif terkait konektivitas dan kemudahan penggunaan.

Wordcloud ini memberikan wawasan bahwa pengguna cenderung memiliki sentimen positif terhadap Starlink, terutama untuk penggunaan di rumah dan paket layanannya. Kata seperti "mau", "coba", dan "pakai" mencerminkan minat pengguna

untuk mencoba layanan ini lebih jauh. Berikut adalah visualisasi kata yang bernilai sentimen negatif dalam bentuk *wordcloud* seperti pada Gambar 3.23.



Gambar 3.23 Visualisasi Kata yang Bernilai Sentimen Negatif dalam Bentuk *Wordcloud*

Untuk *wordcloud* sentimen negatif menampilkan kata – kata yang sering muncul dengan konotasi negatif dalam dataset teks. Kata – kata seperti "Starlink", "sinyal", "Indonesia", "bayar", dan "pakai" terlihat dominan karena ukurannya lebih besar, mencerminkan frekuensi kemunculan yang tinggi. Hal ini menunjukkan bahwa aspek – aspek seperti konektivitas, harga, dan penggunaan layanan menjadi fokus utama kritik atau keluhan.

Wordcloud ini membantu mengidentifikasi topik utama yang menjadi perhatian pengguna, seperti masalah sinyal, harga layanan, dan ketidakpuasan terhadap fitur tertentu. Kata – kata seperti "bukan", "jangan", dan "mahal" juga mencerminkan ketidakpuasan yang berkaitan dengan ekspektasi yang tidak terpenuhi.

Sedangkan, untuk *wordcloud* sentimen netral menampilkan kata – kata yang sering muncul dengan konotasi netral dalam dataset teks seperti pada Gambar 3.24. Kata – kata seperti "starlink", "internet", "pakai", "berapa", "wifi", dan "review" tampak menonjol, menunjukkan bahwa banyak pengguna lebih fokus pada aspek teknis, harga, penggunaan, dan ulasan produk secara umum tanpa menunjukkan sentimen emosional tertentu. Kata seperti "indihome", "listrik", "satelit", dan "daerah" juga muncul cukup sering, mengindikasikan perbandingan dengan penyedia layanan lain serta pembahasan terkait infrastruktur dan lokasi penggunaan.



Gambar 3.24 Visualisasi Kata yang Bernilai Sentimen Netral dalam Bentuk *Wordcloud*

Wordcloud ini memberikan gambaran awal bahwa banyak komentar bersifat informatif atau pertanyaan, dan belum mengandung opini yang kuat, menjadikannya relevan untuk dikategorikan sebagai sentimen netral.

3.4 Rencana Pengujian

Setelah data mendapatkan nilai sentimennya, langkah selanjutnya adalah melakukan proses pengujian model klasifikasi sentimen. Pengujian ini bertujuan

untuk mengevaluasi kinerja model dalam memprediksi sentimen komentar serta menentukan konfigurasi terbaik untuk diterapkan dalam pembuatan model prediksi sederhana. penulis akan menentukan rencana pengujian. Pengujian ini akan dibahas pada bab 4 penelitian ini dan dilakukan melalui beberapa skenario. Berikut adalah skenario pengujian yang akan dilakukan meliputi:

1. Mengacu pada penelitian (Reza et al., 2025) dan (Muslim et al., 2024), model akan diuji dengan berbagai proporsi rasio pembagian data latih dan data uji yang berbeda, yaitu menggunakan rasio 60:40, 70:30, dan 80:20 untuk melihat bagaimana performa model ketika jumlah data latih dan data ujinya berbeda. Pada setiap rasio, algoritma yang digunakan adalah *Random Forest*, dan evaluasi dilakukan menggunakan *confusion matrix* yang berisi akurasi, presisi, recall, dan *F1-score*. Hasil dari skenario ini akan dibandingkan untuk menentukan kombinasi pembagian data terbaik.
2. Setelah memperoleh hasil evaluasi dari berbagai proporsi data latih dan uji, proporsi dengan performa terbaik akan digunakan sebagai dasar dalam membangun model prediksi sentimen sederhana.
3. Model prediksi sederhana yang telah dibangun akan diuji kembali menggunakan beberapa kalimat baru. Kalimat – kalimat ini dipilih secara khusus untuk mewakili variasi gaya bahasa pengguna, seperti bahasa gaul, komentar panjang, hingga komentar yang mengandung unsur sarkasme. Pengujian ini bertujuan untuk mengetahui seberapa baik model dapat memahami konteks dan memprediksi sentimen pada kalimat yang tidak terdapat dalam data pelatihan.

BAB IV
HASIL DAN PEMBAHASAN

4.1 Pengujian Skenario

Pada bab ini, dilakukan pengujian model berdasarkan skenario yang telah direncanakan di bab 3.5. Pengujian skenario ini diawali dengan membagi data latih dan data uji dengan rasio yang berbeda, penerapan algoritna *random forest*, pembuatan model prediksi sentimen, dan dilakukan pengujian menggunakan beberapa komentar baru dari model prediksi sentimen yang telah dibuat. Berikut adalah hasil dari pengujian untuk setiap skenario:

4.1.1 Pembagian Data

Langkah awal dalam proses pembagian data latih dan data uji adalah dengan mengimpor fungsi train_test_split dari *library* sklearn.model_selection, seperti ditunjukkan pada Gambar 4.1. Fungsi ini digunakan untuk membagi dataset menjadi data latih dan data uji secara acak, namun terkontrol melalui *parameter* random_state. Pembagian data latih dan data uji dilakukan menggunakan rasio 60:40, 70:30, dan 80:20.

```
from sklearn.model_selection import train_test_split
```

Gambar 4.1 Import Library untuk Pembagian Data

Fungsi ini digunakan untuk membagi dataset menjadi data latih dan data uji secara acak, namun terkontrol melalui *parameter* random_state. Pembagian data latih dan data uji dilakukan menggunakan rasio 60:40, 70:30, dan 80:20. Pembagian ini dilakukan untuk mengetahui bagaimana algoritma *random forest* berubah ketika proporsi data latih dan data uji memiliki beberapa variasi. Data latih digunakan untuk melatih model, sedangkan data uji digunakan untuk menguji model dalam mengklasifikasikan sentimen. Berikut adalah hasil dari tahapan pembagian data:

A. 60:40

Kode untuk membagi data dengan rasio 60 persen data latih dan 40 persen data uji ditunjukkan pada Gambar 4.2. Dalam implementasinya, fungsi train_test_split menerima parameter test_size=0.4 yang menandakan bahwa 40 persen data akan digunakan sebagai data uji.

```
# Splitting 60:40

x_train_40, x_test_40, y_train_40, y_test_40 =
```

```
train_test_split(x, y, test_size=0.4, random_state=42)
```

Gambar 4.2 Implementasi Kode Python untuk Pembagian Data Latih dan Data Uji dengan Rasio 60:40

Berikut adalah rincian perhitungan data latih dan data uji:

- Data uji = 40% dari 7.384

0.4 × 7.384 = 2.953,6

Dibulatkan menjadi 2.954 data uji

- Data latih = 60% dari 7.384

7.384 – 2.954 = 4.430

Jadi, ada 4.430 data latih

B. 70:30

Proses pembagian data dengan rasio 70:30 dilakukan dengan menyesuaikan parameter `test_size` menjadi 0.3, seperti yang ditampilkan pada Gambar 4.3.

```
# Splitting 70:30

x_train_30, x_test_30, y_train_30, y_test_30 =
train_test_split(x, y, test_size=0.3, random_state=42)
```

Gambar 4.3 Implementasi Kode Python untuk Pembagian Data Latih dan Data Uji dengan Rasio 70:30

Berikut adalah rincian perhitungan data latih dan data uji:

- Data uji = 30% dari 7.384

0.3 × 7.384 = 2.215,2

Dibulatkan menjadi 2.216 data uji

- Data latih = 70% dari 7.384

7.384 – 2.216 = 5.168

Jadi, ada 5.168 data latih

C. 80:20

Konfigurasi terakhir adalah pembagian data dengan rasio 80:20, di mana 20 persen data digunkanan sebagai data uji. Kode yang digunakan untuk skenario ini dapat dilihat pada

Gambar 4.4.

```
# Splitting 80:20

x_train_20, x_test_20, y_train_20, y_test_20 =
train_test_split(x, y, test_size=0.2, random_state=42)
```

Gambar 4.4 Implementasi Kode Python untuk Pembagian Data Latih dan Data Uji dengan Rasio 80:20

Berikut adalah rincian perhitungan data latih dan data uji:

- Data uji = 20% dari 7.384

0.2 × 7.384 = 1.476,8

Dibulatkan menjadi 1.477 data latih

- Data latih = 80% dari 7.384

7.384 – 1.477 = 5.907

Jadi, ada 5.907 data uji

4.1.2 Algoritma

Setelah data dibagi, algoritma *random forest* diterapkan pada setiap rasio pembagian data. Pada setiap rasio, model dilatih menggunakan data latih dan diuji menggunakan data uji. Proses pelatihan melibatkan pembentukan pohon keputusan *(decision trees)* yang bekerja secara *ensemble* untuk meningkatkan akurasi dan mengurangi overfitting. Gambar 4.5 menunjukkan bagaimana model dibuat menggunakan `RandomForestClassifier` dari *library* `sklearn.ensemble`, dengan parameter `n_estimators` ditetapkan sebanyak 200.

```
from sklearn.ensemble import RandomForestClassifier

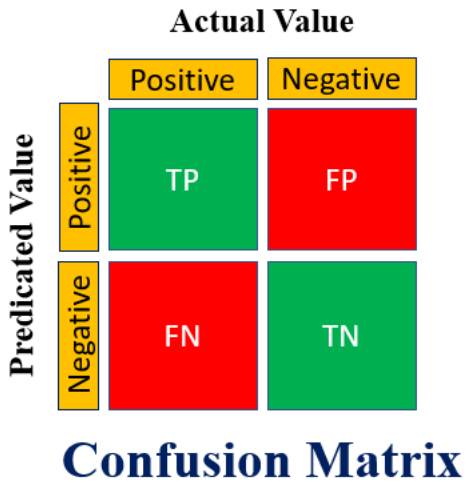
rf = RandomForestClassifier(n_estimators=200, random_state=42)

model = Pipeline([('vectorizer', tvec), ('classifier', rf)])
```

Gambar 4.5 Implementasi Model Random Forest Menggunakan Pipeline dari Scikit-Learn

4.1.3 Evaluasi Performa Model

Evaluasi performa model adalah proses untuk mengukur sejauh mana model yang telah dibangun dapat memprediksi sentimen. Evaluasi model melibatkan *confusion matrix* untuk melihat prediksi yang benar dan salah. *Confusion matrix* adalah alat yang digunakan untuk mengevaluasi performa model klasifikasi dalam *machine learning*. Matriks ini membandingkan prediksi model dengan nilai aktual (ground truth) dan membagi hasilnya menjadi empat kategori, yaitu *True Positive (TP), True Negative (TN), False Positive (FP),* dan *False Negative (FN).* Gambar 4.6 merupakan bentuk dari *confusion matrix.*



Gambar 4.6 Confusion Matrix

Berikut adalah penjelasan masing – masing:

1. *True Positive (TP),* jumlah sampel yang diprediksi positif oleh model dan memang benar positif sesuai data aktual.
2. *True Negative (TN),* jumlah sampel yang diprediksi negatif oleh model dan memang benar negatif sesuai data aktual.
3. *False Positive (FP)* jumlah sampel yang diprediksi positif oleh model, tetapi sebenarnya negatif
4. *False Negative (FN)* jumlah sampel yang diprediksi negatif oleh model, tetapi sebenarnya positif

Dari *confusion matrix*, dapat dihitung nilai *accuracy*, *precision*, *recall*, dan *F1-score* dari sebuah model yang telah dibuat. Setiap model yang telah dilatih dievaluasi menggunakan metrik – metrik seperti *accuracy*, *precision*, *recall*, dan *F1-score*. *Accuracy* atau akurasi digunakan untuk mengukur seberapa sering model melakukan prediksi yang benar, sementara *precision* dan *recall* digunakan untuk mengevaluasi performa model pada setiap kelas sentimen (positif, negatif, dan netral). Sementara itu, *F1-score* digunakan sebagai rata – rata harmonik

antara *precision* dan *recall* untuk memberikan gambaran yang lebih seimbang tentang performa model. Rasio pembagian data yang menghasilkan performa terbaik (misalnya, akurasi tertinggi) dipilih untuk langkah selanjutnya

Adapun rumus untuk menghitung nilai *accuracy*, *precision*, *recall*, dan *F1-score* sebagai berikut:

a. *Accuracy*

$$Accuracy = \frac{\text{Jumlah prediksi benar}}{\text{Total data}} \tag{4.1}$$

b. *Precision*

$$Precison = \frac{True\ Positive\ (TP)}{True\ Positive\ (TP) + False\ Positive\ (FP)} \tag{4.2}$$

c. *Recall (Sensivity* atau *True Positive Rate)*

$$Recall = \frac{True\ Positive\ (TP)}{True\ Positive\ (TP) + False\ Negative\ (FN)} \tag{4.3}$$

d. *F1-score*

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precison + Recall} \tag{4.4}$$

Berikut adalah *confusion matrix* dan perhitungan nilai – nilai *accuracy, precision, recall* dan *F1-score* pada setiap rasio:

a. 40:60

Gambar 4.7 menunjukkan hasil evaluasi model random forest menggunakan rasio 60:40, sedangkan Gambar 4.8 adalah bentuk dari confusion matrix-nya.

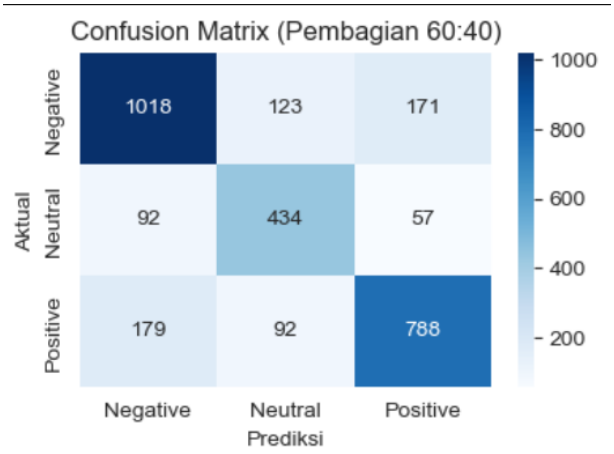
```
Evaluasi Model (Random Forest) - Pembagian 60:40
================================================
Akurasi: 0.7583
Recall: 0.7583
Presisi: 0.7608
F1-Score: 0.7590

Laporan Klasifikasi:
              precision    recall  f1-score   support

    Negative       0.79      0.78      0.78      1312
     Neutral       0.67      0.74      0.70       583
    Positive       0.78      0.74      0.76      1059

    accuracy                           0.76      2954
   macro avg       0.74      0.75      0.75      2954
weighted avg       0.76      0.76      0.76      2954
```

Gambar 4.7 Hasil Evaluasi Model pada Rasio 60:40



Gambar 4.8 Confusion Matrix pada Rasio 60:40

1. *Accuracy*

Dari confusion matrix, jumlah prediksi benar adalah:

- True Positive untuk kelas Negative: 1018
- True Positive untuk kelas Neutral: 434
- True Positive untuk kelas Positive: 788

Total prediksi benar = 1018 + 434 + 788 = 2240

Total data = 2954

Berikut perhitungan nilai akurasi:

$$Accuracy = \frac{2240}{2954} \approx 0.758$$

2. *Precision* (Presisi)

Perhitungan nilai presisi untuk tiap kelas sentimen:

- Kelas Negative

$$Presisi\ Negative = \frac{1018}{1018 + 92 + 179} = \frac{1018}{1289} \approx 0.79$$

- Kelas Positif

$$Presisi\ Positive = \frac{788}{171 + 57 + 788} = \frac{788}{1016} \approx 0.78$$

- Kelas Netral

$$Presisi\ Neutral = \frac{434}{123 + 434 + 92} = \frac{434}{649} \approx 0.67$$

3. *Recall*

Berikut adalah perhitungan nilai *recall* untuk tiap kelas sentimen:

- Kelas Negatif

$$Recall\ Negative = \frac{1018}{1018 + 123 + 171} = \frac{1018}{1312} \approx 0.78$$

- Kelas Positif

$$Recall\ Positive = \frac{788}{179 + 92 + 788} = \frac{788}{1059} \approx 0.74$$

- Kelas Netral

$$Recall\ Neutral = \frac{434}{92 + 434 + 57} = \frac{434}{583} \approx 0.74$$

4. *F1-score*

Berikut adalah perhitungan nilai *F1-score* untuk tiap kelas sentimen:

- Kelas Negatif

$$F1\ Negative = 2 \times \frac{0.79 \times 0.78}{0.79 + 0.78} \approx 0.785$$

- Kelas Positif

$$F1\ Positive = 2 \times \frac{0.78 \times 0.74}{0.78 + 0.74} \approx 0.759$$

- Kelas Netral

$$F1\ Neutral = 2 \times \frac{0.67 \times 0.74}{0.67 + 0.74} \approx 0.703$$

Kesimpulan: Model dengan rasio 40:60 menghasilkan akurasi sebesar 75.8%, menunjukkan bahwa 75,8% dari total prediksi benar. Nilai presisi untuk kelas negatif dan positif cukup tinggi yang dimana masing – masing 0.79 dan 0.78, artinya model cukup baik dalam meminimalkan *false positive* untuk kedua kelas tersebut. Namun, presisi kelas netral (0.67) lebih rendah, yang berarti terdapat lebih banyak prediksi salah untuk kelas ini. *recall* kelas netral (0.74) lebih baik daripada presisinya, menunjukkan bahwa model bisa menangkap sebagian besar kasus nettral, tetapi masih sering salah mengklasifikasikan kelas lain sebagai nettral. Nilai *F1-score* yang seimbang (sekitar 0.76) mengindikasikan bahwa model memiliki trade-off yang baik antara presisi dan recall secara keseluruhan, meskipun performa kelas minor (Netral) masih perlu

ditingkatkan.

b. 70:30

Gambar 4.9 menunjukkan hasil evaluasi model random forest menggunakan rasio 70:23, sedangkan Gambar 4.10 adalah bentuk dari confusion matrix-nya.

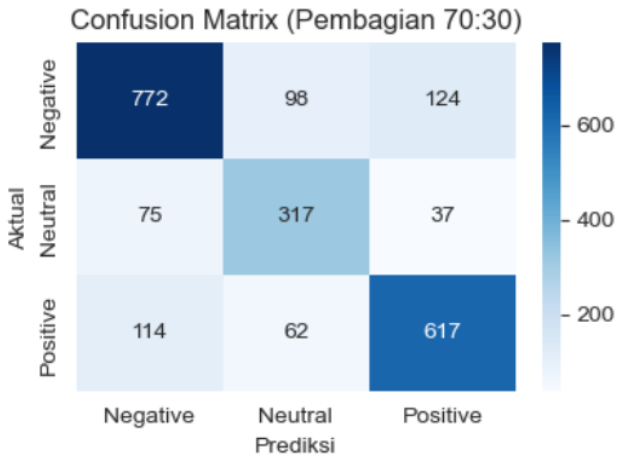
```
Evaluasi Model (Random Forest) - Pembagian 30:70
==================================
Akurasi: 0.7699
Recall: 0.7699
Presisi: 0.7728
F1-Score: 0.7708

Laporan Klasifikasi:
              precision    recall  f1-score   support

    Negative       0.80      0.78      0.79       994
     Neutral       0.66      0.74      0.70       429
    Positive       0.79      0.78      0.79       793

    accuracy                           0.77      2216
   macro avg       0.75      0.76      0.76      2216
weighted avg       0.77      0.77      0.77      2216
```

Gambar 4.9 Hasil Evaluasi Model pada Rasio 70:30



Gambar 4.10 Confusion Matrix pada Rasio 70:30

1. *Accuracy*

Dari confusion matrix, jumlah prediksi benar adalah:

- True Positive untuk kelas Negative: 772
- True Positive untuk kelas Neutral: 317
- True Positive untuk kelas Positive: 617

Total prediksi benar = 772 + 317 + 617 = 1706

Total data = 2216

Berikut perhitungan nilai akurasi:

$$Accuracy = \frac{1706}{2216} \approx 0.77$$

2. *Precision* (Presisi)

Perhitungan nilai presisi untuk tiap kelas sentimen:

- Kelas Negative

$$Presisi\ Negative = \frac{772}{772 + 75 + 114} = \frac{772}{961} \approx 0.80$$

- Kelas Positif

$$Presisi\ Positive = \frac{617}{124 + 37 + 617} = \frac{617}{778} \approx 0.79$$

- Kelas Netral

$$Presisi\ Neutral = \frac{317}{98 + 317 + 62} = \frac{317}{477} \approx 0.66$$

3. *Recall*

Berikut adalah perhitungan nilai *recall* untuk tiap kelas sentimen:

- Kelas Negatif

$$Recall\ Negative = \frac{772}{772 + 98 + 124} = \frac{772}{994} \approx 0.78$$

- Kelas Positif

$$Recall\ Positive = \frac{617}{114 + 62 + 617} = \frac{617}{793} \approx 0.78$$

- Kelas Netral

$$Recall\ Neutral = \frac{317}{75 + 317 + 37} = \frac{317}{429} \approx 0.74$$

4. *F1-score*

Berikut adalah perhitungan nilai *F1-score* untuk tiap kelas sentimen:

- Kelas Negatif

$$F1\ Negative = 2\ \times \frac{0.80 \times 0.78}{0.80 + 0.78} \approx 0.79$$

- Kelas Positif

$$F1\ Positive = 2\ \times \frac{0.79 \times 0.78}{0.79 + 0.78} \approx 0.785$$

- Kelas Netral

$$F1\ Neutral = 2\ \times \frac{0.66 \times 0.74}{0.66 + 0.74} \approx 0.70$$

c. 80:20

Gambar 4.11 menunjukkan hasil evaluasi model *random forest* menggunakan rasio 80:20, sedangkan Gambar 4.12 adalah bentuk dari *confusion matrix*-nya.

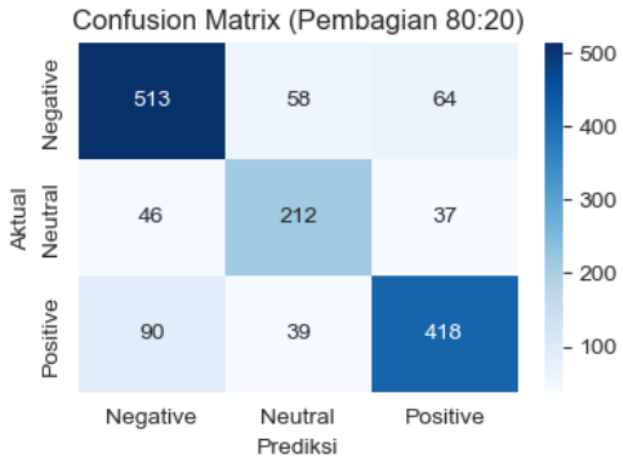
```
Evaluasi Model (Random Forest) - Pembagian 20:80
==================================
Akurasi: 0.7739
Recall: 0.7739
Presisi: 0.7751
F1-Score: 0.7742

Laporan Klasifikasi:
              precision    recall  f1-score   support

    Negative       0.79      0.81      0.80       635
     Neutral       0.69      0.72      0.70       295
    Positive       0.81      0.76      0.78       547

    accuracy                           0.77      1477
   macro avg       0.76      0.76      0.76      1477
weighted avg       0.78      0.77      0.77      1477
```

Gambar 4.11 Hasil Evaluasi Model pada Rasio 80:20



Gambar 4.12 Confusion Matrix pada Rasio 80:20

1. *Accuracy*

Dari confusion matrix, jumlah prediksi benar adalah:

- True Positive untuk kelas Negative: 513
- True Positive untuk kelas Neutral: 212
- True Positive untuk kelas Positive: 418

Total prediksi benar = 513 + 212 + 418 = 1143

Total data = 1477

Berikut perhitungan nilai akurasi:

$$Accuracy = \frac{1143}{1477} \approx 0.774$$

2. *Precision* (Presisi)

Perhitungan nilai presisi untuk tiap kelas sentimen:

- Kelas Negatif

$$Presisi\ Negative = \frac{513}{513 + 46 + 90} = \frac{513}{649} \approx 0.79$$

- Kelas Positif

$$Presisi\ Positive = \frac{418}{64 + 37 + 418} = \frac{418}{519} \approx 0.81$$

- Kelas Netral

$$Presisi\ Neutral = \frac{212}{58 + 212 + 39} = \frac{212}{477} \approx 0.69$$

3. *Recall*

Berikut adalah perhitungan nilai *recall* untuk tiap kelas sentimen:

- Kelas Negatif

$$Recall\ Negative = \frac{513}{513 + 58 + 64} = \frac{513}{635} \approx 0.81$$

- Kelas Positif

$$Recall\ Positive = \frac{418}{90 + 39 + 418} = \frac{418}{547} \approx 0.76$$

- Kelas Netral

$$Recall\ Neutral = \frac{212}{46 + 212 + 37} = \frac{212}{295} \approx 0.72$$

4. *F1-score*

Berikut adalah perhitungan nilai *F1-score* untuk tiap kelas sentimen:

- Kelas Negatif

$$F1\ Negative = 2 \times \frac{0.79 \times 0.81}{0.79 + 0.81} \approx 0.80$$

- Kelas Positif

$$F1\ Positive = 2 \times \frac{0.81 \times 0.76}{0.81 + 0.76} \approx 0.78$$

- Kelas Netral

$$F1\ Neutral = 2 \times \frac{0.81 \times 0.76}{0.81 + 0.76} \approx 0.70$$

Perbandingan performa ketiga model dapat dilihat pada Tabel 4.1.

Tabel 4.1 Performa Ketiga Model

Metrik	**60:40**	**70:30**	**80:20**	**Kesimpulan**
Accuracy	0.758	0.77	0.774	Rasio 80:20 memiliki akurasi paling baik
Precision (weighted average)	0.76	0.77	0.78	Rasio 80:20 lebih konsisten
Recall (weighted average)	0.76	0.77	0.77	Sama, namun rasio 80:20 lebih unggul

F1-Score (weighted average)	0.76	0.77	0.77	Rasio 80:20 lebih stabil

Berdasarkan tabel perbandingan performa ketiga model dengan rasio pembagiann data yang berbeda, dapat dijelaskan bahwa model dengan rasio pembagian 80:20 menunjukkan performa yang paling optimal. Model ini mencapai akurasi sebesar 0.774 (77.4%), mengungguli model dengan rasio 70:30 dan 60:40. Tingginya nilai akurasi pada model 20:80 dapat diartikan bahwa peningkatan proporsi data latih hingga 80% memberikan dampak signifikan terhadap kemampuan model dalam memprediksi data uji.

Dalam metrik presisi yang dihitung menggunakan *weighted average,* model 80:20 unggul denngan nilai 0,78 lebih tinggi dibandingkan model 70:30 dengan nilai 0,77 dan model 60:40 dengan nilai 0,76. Nilai presisi yang tinggi ini menunjukkan konsistensi model dalam meminimalkan kesalahan klasifikasi *false positive,* khususnya untuk kelas minor seperti sentimen netral. Hal ini menjadi aspek penting mengingat tantangan utama dalam klasifikasi sentimen adalah kemampuan model dalam menangani ketidakseimbangan kelas.

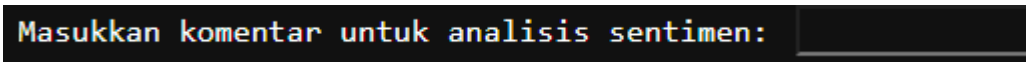
Pada metrik *recall* dengan perhitungan *weighted average,* ketika model menunukkan performa yang relative setara dengan nilai 0,76 – 0,77. Namun, model 80:20 sebagai nilai terbaik dengan nilai *recall* sebesar 0,77 terutama didukung oleh kemampuan identifikasi yang sangat baik untuk kelas negative (*recall* 0,81). Kemampuan model dalam mencapai nilai *recall* yang tinggi untuk kelas dominan ini menjadi keunggulan yang patut diperhitungkan.

F1-score sebagai matrik yang memadukan presisi dan *recall* juga menempatkan model 80:20 sebagai yang terbaik dengan nilai 0,77. Kestabilan nilai *F1-score* pada model ini menunjukkan keseimbangan optimal antara kemampuan meminimalkan kesalahan klasifikasi (presisi) dan memaksimalkan indentifikasi kasus yang benar *(recall).* Performa yang konsisten ini menjadikan 80:20 sebagai pilihan yang lebih unggul.

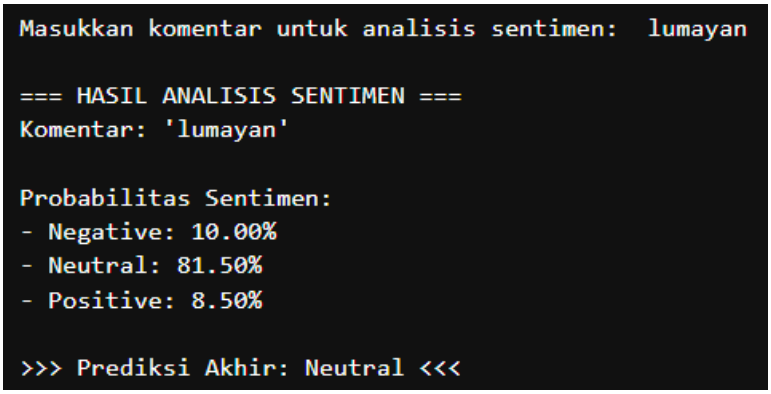
4.1.4 Modelling

Pada tahap ini, model prediksi sentimen dibuat menggunakan dasar pembagian rasio data 80:20 dan algoritma *random forest.* Model dibuat dalam bentuk *pipeline* yang menerima *input* berupa kalimat dari pengguna, lalu memproses kalimat tersebut melalui tahap vektorisasi dan

klasifikasi, sebelum akhirnya menghasilkan prediksi sentimen. Gambar 4.13 menunjukkan model prediksi sentimen dan Gambar 4.14 adalah hasil *output* model.



Gambar 4.13 Model Prediksi Sentimen



Gambar 4.14 Output dari Model Prediksi Sentimen

4.1.5 Testing

Selanjutnya, model akan diujikan dengan beberapa kalimat yang mengandung sentimen positif, negatif, dan netral, serta beberapa kalimat yang mengandung unsur sarkasme dan juga kalimat panjang baru yang belum pernah digunakan dalam proses pelatihan maupun pengujian sebelumnya. Tujuan dari pengujian ini adalah untuk menilai kemampuan model dalam mengklasifikasikan sentimen secara langsung berdasarkan *input* teks yang belum dikenali sebelumnya. Daftar kalimat yang akan digunakan dalam proses pengujian disajikan pada Tabel 4.2

Tabel 4.2 Daftar Kalimat untuk Pengujian Model

No.	**Kalimat**
1.	“Starlink ngebantu banget buat sekolah *online* anak – anak di daerah *gue*.”
2.	“Akhirnya bisa kerja *remote* dari desa tanpa drama sinyal, makasih Starlink!”
3.	“Pernah nyoba Starlink seminggu, tapi *buffering* terus padahal sinyal *full*, *zonk* banget.”
4.	“Internet dari Starlink sangat cepat dan stabil, cocok banget untuk daerah terpencil di Indonesia yang sebelumnya sulit akses internet.”

5.	"*Lo* bisa order Starlink langsung *via web* – nya, tinggal nunggu kiriman alatnya."
6.	"Banyak daerah di Indo udah mulai bisa akses Starlink, tergantung wilayah sih."
7.	"Bayar jutaan per bulan biar bisa nikmatin sinyal ilang – timbul? *Worth* it banget lah ya!"
8.	"Udah mahal, pemasangannya ribet, eh internetnya juga nggak selalu jalan — *top* lah!"
9.	"Sinyalnya stabil, bisa *video call* tanpa putus, *recommended* banget!"
10.	"Langit cerah, tapi sinyal ilang, ini sih *ngaco*."
11.	"Kenceng sih… tapi kenceng ilangnya juga. Mantap *bet* dah."
12.	"Starlink tuh internet satelit punya SpaceX, *bro*."
13.	"Lapor ke *CS* kayak ngomong sama tembok, no respon men."
14.	"Baru nyobain, langsung jatuh cinta ama kecepatannya."
15.	"Makasih loh Starlink, sekarang *buffering* jadi agenda keluarga tiap malam."
16.	"*Gue* udah pasang Starlink di rumah nenek yang ada di pelosok, dan walaupun awalnya sempet ragu karena takut sinyalnya jelek, ternyata setelah dipasang dan dicoba buat streaming *YouTube* sama main *game online*, hasilnya lumayan stabil dan bisa diandalkan, meskipun kadang – kadang suka ngadat pas hujan deras."
17.	"Banyak yang bilang Starlink itu solusi banget buat pelosok, dan *gue* setuju sih, tapi tetep aja ya, kalau ada halangan kayak pohon *gede* atau cuaca jelek, sinyalnya bisa ilang – ilangan, jadi tetep mesti realistis lah ekspektasinya."
18.	"Baru beberapa minggu pake Starlink, udah ngerasain sinyal ilang sendiri tengah malem, terus pas mau *reconnect* susah banget, padahal katanya koneksi satelit lebih stabil dari *BTS* biasa."

Setelah daftar kalimat uji disusun, model kemudian digunakan untuk mengklasifikasikan sentimen dari masing – masing kalimat tersebut. Proses ini tidak hanya menghasilkan prediksi sentimen akhir, tetapi juga menampilkan probabilitas untuk setiap kategori sentiment, sehingga dapat memberikan gambaran yang lebih mendalam mengenai tingkat keyakinan model

terhadap setiap prediksi yang dihasilkan. Untuk hasil dari pengujian dapat dilihat pada tabel 4.3.

Tabel 4.3 Daftar Kalimat dan Hasil Pengujian Model

No.	Kalimat	Label Sebenarnya	Label Model	Prediksi Model
1.	"Starlink ngebantu banget buat sekolah online anak – anak di daerah gue."	Positif	Positif	Probabilitas Sentimen: - Negative: 31.50% - Neutral: 8.50% - Positive: 60.00% Prediksi Akhir: Positive
2.	"Akhirnya bisa kerja *remote* dari desa tanpa drama sinyal, makasih Starlink!"	Positif	Positif	Probabilitas Sentimen: - Negative: 31.50% - Neutral: 23.00% - Positive: 45.50% Prediksi Akhir: Positive
3.	"Pernah nyoba Starlink seminggu, tapi *buffering* terus padahal sinyal full, zonk banget."	Negatif	Negatif	Probabilitas Sentimen: - Negative: 46.50% - Neutral: 20.00% - Positive: 33.50% Prediksi Akhir: Negative
4.	"Internet dari Starlink sangat cepat dan stabil, cocok banget untuk daerah terpencil di Indonesia yang sebelumnya sulit akses internet."	Positif	Positif	Probabilitas Sentimen: - Negative: 29.50% - Neutral: 5.50% - Positive: 65.00% Prediksi Akhir: Positive
5.	"Lo bisa order Starlink langsung *via web*–nya, tinggal nunggu kiriman alatnya."	Netral	Positif	Probabilitas Sentimen: - Negative: 35.50% - Neutral: 13.00% - Positive: 51.50% Prediksi Akhir: Positive
6.	"Banyak daerah di Indo udah mulai bisa akses Starlink, tergantung wilayah sih."	Netral	Negatif	Probabilitas Sentimen: - Negative: 55.00% - Neutral: 16.50% - Positive: 28.50% Prediksi Akhir: Negative
7.	"Bayar jutaan per bulan biar bisa nikmatin sinyal ilang – timbul? *Worth it* banget lah ya!"	Sarkasme, negatif	Negatif	Probabilitas Sentimen: - Negative: 58.50% - Neutral: 9.00% - Positive: 32.50%

				Prediksi Akhir: Negative
8.	"Udah mahal, pemasangannya ribet, eh internetnya juga nggak selalu jalan, *top* lah!"	Sarkasme, negatif	Positif	Probabilitas Sentimen: - Negative: 6.50% - Neutral: 5.50% - Positive: 88.00% Prediksi Akhir: Positive
9.	"Sinyalnya stabil, bisa *video call* tanpa putus, *recommended* banget!"	Positif	Positif	Probabilitas Sentimen: - Negative: 40.00% - Neutral: 15.50% - Positive: 44.50% Prediksi Akhir: Positive
10.	"Langit cerah, tapi sinyal ilang, ini sih *ngaco*."	Negatif	Netral	Probabilitas Sentimen: - Negative: 36.50% - Neutral: 48.00% - Positive: 15.50% Prediksi Akhir: Neutral
11.	"Kenceng sih… tapi kenceng ilangnya juga. Mantap bet dah."	Sarkasme negatif	Positif	Probabilitas Sentimen: - Negative: 28.00% - Neutral: 1.00% - Positive: 71.00% Prediksi Akhir: Positive
12.	"Starlink tuh internet satelit punya *SpaceX*, bro."	Netral	Negatif	Probabilitas Sentimen: - Negative: 72.50% - Neutral: 8.50% - Positive: 19.00% Prediksi Akhir: Negative
13.	"Lapor ke *CS* kayak ngomong sama tembok, no respon men."	Negatif	Positif	Probabilitas Sentimen: - Negative: 14.00% - Neutral: 0.00% - Positive: 86.00% Prediksi Akhir: Positive
14.	"Baru nyobain, langsung jatuh cinta ama kecepatannya."	Positif	Positif	Probabilitas Sentimen: - Negative: 18.00% - Neutral: 9.50% - Positive: 72.50% Prediksi Akhir: Positive
15.	"Makasih loh Starlink, sekarang *buffering* jadi agenda keluarga tiap malam."	Sarkasme, negatif	Negatif	Probabilitas Sentimen: - Negative: 50.50% - Neutral: 17.50% - Positive: 32.00%

				Prediksi Akhir: Negative
16.	"Gue udah pasang Starlink di rumah nenek yang ada di pelosok, dan walaupun awalnya sempet ragu karena takut sinyalnya jelek, ternyata setelah dipasang dan dicoba buat streaming YouTube sama main game online, hasilnya lumayan stabil dan bisa diandalkan, meskipun kadang-kadang suka ngadat pas hujan deras."	Positif	Positif	Probabilitas Sentimen: - Negative: 36.50% - Neutral: 2.50% - Positive: 61.00% Prediksi Akhir: Positive
17.	"Banyak yang bilang Starlink itu solusi banget buat pelosok, dan gue setuju sih, tapi tetep aja ya, kalau ada halangan kayak pohon gede atau cuaca jelek, sinyalnya bisa ilang-ilangan, jadi tetep mesti realistis lah ekspektasinya."	Netral	Negatif	Probabilitas Sentimen: - Negative: 69.00% - Neutral: 6.50% - Positive: 24.50% Prediksi Akhir: Negative
18.	"Baru beberapa minggu pake Starlink, udah ngerasain sinyal ilang sendiri tengah malem, terus pas mau reconnect susah banget, padahal katanya koneksi satelit lebih stabil dari BTS biasa."	Negatif	Negatif	Probabilitas Sentimen: - Negative: 63.50% - Neutral: 2.50% - Positive: 34.00% Prediksi Akhir: Negative

Berdasarkan hasil pengujian yang telah dilakukan, dapat disimpulkan bahwa model mampu mengklasifikasikan sebagian besar sentimen dengan cukup baik, khususnya pada kategori sentimen positif dan negatif. Model menunjukkan performa yang cukup konsisten dalam mendeteksi kalimat – kalimat dengan ekspresi yang jelas, baik berupa pujian maupun keluhan. Secara umum, model cukup akurat dalam mengidentifikasi kalimat dengan sentimen positif dan negatif.

Namun, beberapa kalimat yang sebenarnya bersifat netral justru diprediksi sebagai positif atau negatif. Hal ini terutama terjadi pada kalimat yang berisi informasi faktual atau deskriptif tanpa muatan emosional yang kuat, di mana model cenderung salah mengidentifikasi sentimen berdasarkan kata – kata yang lebih dominan muncul dalam kalimat. Misalnya, kalimat yang menyatakan kondisi atau fakta tentang produk dapat dianggap positif atau negatif oleh model meskipun tidak ada ekspresi opini yang jelas.

Selain itu, kalimat dengan unsur sarkasme juga sering menjadi tantangan besar bagi model ini. Sarkasme, yang seringkali berlawanan dengan makna harfiah kalimat, mengakibatkan model salah dalam menentukan sentimen. Kalimat yang seharusnya diprediksi sebagai negatif, sering kali diprediksi positif karena model gagal menangkap konotasi sarkastik yang terkandung dalam kalimat tersebut.

BAB V
KESIMPULAN DAN SARAN

5.1 Kesimpulan

Berdasarkan hasil penelitian analisis sentimen terhadap 7,384 komentar YouTube terkait layanan Starlink di Indonesia menggunakan algoritma *Random Forest* dan pendekatan *lexicon-dictionary,* diperolah beberapa kesimpulan sebagai berikut:

1. Penelitian berhasil menggali opini public dari komentar YouTube dengan menerapkan analisis sentiment berbasis *lexicon.* Komentar – komentar dikategorikan menjadi sentimen positif, negatif, dan netral sehingga memberikan gambaran persepsi masyarakat terhadap layanan Starlink.

2. Persepsi publik terhadap layanan Starlink:

    - Analisis sentimen terhadap 7.384 komentar dari video YouTube kanal GadgetIn dan NanangMrk menunjukkan distribusi sentimen sebagai berikut:

        - Positif: 2.627 komentar (35,6%)
        - Negatif: 3.296 komentar (44,6%)
        - Netral: 1.461 komentar (19,8%).

    - Kata – kata dominan dalam sentimen positif antara lain "stabil", "cepat", dan "rekomendasi", sedangkan sentimen negatif didominasi oleh kata seperti "mahal", "sinyal", dan "*buffering*".

3. Pembangunan model prediksi sentimen:

    Model prediksi sentimen berhasil dibuat mengunakan algoritma *Random Forest,* dengan performa terbaik pada rasio 80:20. Menghasilkan performa terbaik dengan akurasi 77,4%, presisi 78%, recall 77%, dan F1-score 77%.

4. Kemampuan *Random Forest* dalam mengotomatisasi klasifikasi dan prediksi sentimen:

    Model *Random Forest* mampu mengotomatisasi klasifikasi sentimen dengan performa yang konsisten, terutama dalam mengenali komentar positif dan negatif. Namun, terdapat tantangan dalam menangani komentar netral dan

kalimat bermuatan sarkasme, yang seringkali salah klasifikasi. Hal ini menjadi catatan penting bagi pengembangan model di masa depan.

5. Implikasi Penelitian:
    - Hasil analisis mencerminkan polarisasi opini netizen terhadap Starlink, dengan mayoritas komentar bernada negatif terkait harga dan kestabilan koneksi.
    - Keberhasilan model *Random Forest* dalam analisis ini mendukung potensi penerapannya untuk klasifikasi sentimen berbasis teks berbahasa Indonesia.

5.2 Saran

Berdasarkan temuan dan keterbatasan penelitian ini, berikut rekomendasi untuk pengembangan lebih lanjut:

1. Peningkatan Model

   Menggunakan algoritma lain seperti BERT atau LSTM untuk menangani konteks kalimat kompleks (termasuk sarkasme) dan meningkatkan akurasi klasifikasi sentimen netral.
2. Pengembangan Dataset

   Memperbanyak data latih dengan komentar dari platform lain (Twitter, Instagram) untuk meningkatkan generalisasi model. Menyeimbangkan distribusi kelas sentimen (positif, negatif, netral) agar model tidak bias terhadap kelas mayoritas.
3. Aplikasi Praktis

   Model dapat diintegrasikan dengan sistem social media monitoring untuk perusahaan seperti SpaceX atau penyedia layanan internet guna mengevaluasi feedback pengguna secara *real-time.* Penelitian serupa dapat diterapkan pada topik teknologi lainnya dengan menyesuaikan lexicon dan dataset.
4. Validasi Tambahan

   Melakukan uji coba model dengan sampel yang lebih besar dan beragam untuk memvalidasi stablitas performa. Mempertimbangkan analisis *berbasis aspect-based sentiment analysis (ABSA)* untuk mengidentifikasi sentimen spesifik terhadap fitur tertentu (misalnya: harga, kecepatan, layanan).

5.3 Penutup

Penelitian ini telah berhasil menganalisis sentimen netizen terhadap Starlink di Indonesia menggunakan algoritma Random Forest, sekaligus menyoroti potensi dan tantangan dalam pemrosesan teks berbahasa Indonesia. Hasilnya dapat menjadi dasar untuk pengembangan sistem analisis sentimen yang lebih canggih, serta masukan bagi pihak terkait dalam mengevaluasi layanan berbasis teknologi.

DAFTAR PUSTAKA

Aftab, F., Bazai, S. U., Marjan, S., Baloch, L., Aslam, S., Amphawan, A., & Neo, T. K. (2023). A Comprehensive Survey on Sentiment Analysis Techniques. *International Journal of Technology*, *14*(6), 1288–1298. https://doi.org/10.14716/ijtech.v14i6.6632

Ahmad, A., & Gata, W. (2022). Sentimen Analisis Masyarakat Indonesia di Twitter Terkait Metaverse dengan Algoritma Support Vector Machine. *Jurnal Teknologi Informasi Dan Komunikasi)*, *6*(4), 2022. https://doi.org/10.35870/jti

Ainapure, B. S., Pise, R. N., Reddy, P., Appasani, B., Srinivasulu, A., Khan, M. S., & Bizon, N. (2023). Sentiment Analysis of COVID-19 Tweets Using Deep Learning and Lexicon-Based Approaches. *Sustainability (Switzerland)*, *15*(3). https://doi.org/10.3390/su15032573

Ajijah, N., & Kurniawan, A. (2023). Klasifikasi Teks Mining Terhadap Analisa Isu Kegiatan Tenaga Lapangan Menggunakan Algoritma K-Nearest Neighbor (KNN). In *Jurnal Sains Komputer & Informatika (J-SAKTI* (Vol. 7, Issue 1).

Akmala, S. (2019). PERKEMBANGAN INTERNET PADA GENERASI MUDA DI INDONESIA DENGAN KAITAN UNDANG-UNDANG ITE YANG BERLAKU. *Cyber Security Dan Forensik Digital*, *1*(2). https://doi.org/10.14421/csecurity.2018.1.2.1371

Apriliani, D., Susanto, A., Fikri Hidayattullah, M., & Wiro Sasmito, G. (2023). *Sentimen Analisis Pandangan Masyarakat Terhadap Vaksinasi Covid 19 Menggunakan K-Nearest Neighbors*. *8*(1).

Azhari, M., Situmorang, Z., & Rosnelly, R. (2021). Perbandingan Akurasi, Recall, dan Presisi Klasifikasi pada Algoritma C4.5, Random Forest, SVM dan Naive Bayes. *JURNAL MEDIA INFORMATIKA BUDIDARMA*, *5*(2), 640. https://doi.org/10.30865/mib.v5i2.2937

Catelli, R., Pelosi, S., & Esposito, M. (2022). Lexicon-Based vs. Bert-Based Sentiment Analysis: A Comparative Study in Italian. *Electronics (Switzerland)*, *11*(3). https://doi.org/10.3390/electronics11030374

Cortis, K., & Davis, B. (2021). Over a decade of social opinion mining: a systematic review. *Artificial Intelligence Review*, *54*(7), 4873–4965. https://doi.org/10.1007/s10462-021-10030-2

Dewi, A. R., Diana, S., Fakhrezi, M. A., Awang, N., Ma'arif, H., & Saputra, D. D. (2023). SENTIMEN ANALISIS TERHADAP PUAN MAHARANI SEBAGAI KANDIDAT

CALON PRESIDEN 2024 BERDASARKAN OPINI TWITTER MENGGUNAKAN METODE NAIVE BAYES DAN ADABOOST. *JSiI (Jurnal Sistem Informasi)*, *10*(1), 75–80. https://doi.org/10.30656/jsii.v10i1.5785

Dewi, S., & Arianto, D. B. (n.d.). *TWITTER SENTIMENT ANALYSIS TOWARDS QATAR AS HOST OF THE 2022 WORLD CUP USING TEXTBLOB*. http://ijsr.internationaljournallabs.com/index.php/ijsr

Epoka, B. E. (2023). Literature Review of Qualitative Data with Natural Language Processing. *Journal of Robotics Spectrum*, 56–65. https://doi.org/10.53759/9852/jrs202301006

Fatma Sjoraida, D., Wibawa, B., Guna, K., & Yudhakusuma, D. (2024). Analisis Sentimen Film Dirty Vote Menggunakan BERT (Bidirectional Encoder Representations from Transformers). *Universitas Langlangbuana*, *8*(2). https://doi.org/10.35870/jti

Fudholi, L. A., Rahaningsih, N., & Dana, R. D. (2024). SENTIMEN ANALISIS PERILAKU PENGGEMAR COLDPLAY DI MEDIA SOSIAL TWITTER MENGGUNAKAN METODE NAIVE BAYES. In *Jurnal Mahasiswa Teknik Informatika* (Vol. 8, Issue 3).

Hanafiah, A., Haza Nasution, A., Arta, Y., Wandri, R., Nasution, H. O., Mardafora, J., & Riau, U. I. (2023). SENTIMEN ANALISIS TERHADAP CUSTOMER REVIEW PRODUK SHOPEE BERBASIS WORDCLOUD DENGAN ALGORITMA NAÏVE BAYES CLASSIFIER SENTIMENT ANALYSIS OF CUSTOMER REVIEWS OF SHOPEE PRODUCTS BASED ON WORDCLOUD USING NAÏVE BAYES CLASSIFIER ALGORITHM. *Journal of Information Technology and Computer Science (INTECOMS)*, *6*(1).

Hassani, H., Beneki, C., Unger, S., Mazinani, M. T., & Yeganegi, M. R. (2020). Text mining in big data analytics. *Big Data and Cognitive Computing*, *4*(1), 1–34. https://doi.org/10.3390/bdcc4010001

Hayaty, M., Adi, S., & Hartanto, A. D. (2020). Lexicon-Based Indonesian Local Language Abusive Words Dictionary to Detect Hate Speech in Social Media. *Journal of Information Systems Engineering and Business Intelligence*, *6*(1), 9. https://doi.org/10.20473/jisebi.6.1.9-17

Herjanto, M. F. Y., & Carudin, C. (2024). ANALISIS SENTIMEN ULASAN PENGGUNA APLIKASI SIREKAP PADA PLAY STORE MENGGUNAKAN ALGORITMA RANDOM FOREST CLASSIFER. *Jurnal Informatika Dan Teknik Elektro Terapan*, *12*(2). https://doi.org/10.23960/jitet.v12i2.4192

Imaduddin, H., Yusfida A'la, F., & Nugroho, Y. S. (n.d.). Sentiment Analysis in Indonesian Healthcare Applications using IndoBERT Approach. In *IJACSA) International Journal of Advanced Computer Science and Applications* (Vol. 14, Issue 8). www.ijacsa.thesai.org

Iqbal, M., & Yusuf, R. (n.d.). *VISUALISASI KATA KUNCI PEMBERITAAN PEMILU 2024 MENGGUNAKAN SPACY DAN WORDCLOUD20240617*.

Kaburuan, E. R., & Setiawan, N. R. (2023). Sentimen Analisis Review Aplikasi Digital Korlantas Pada Google Play Store Menggunakan Metode SVM. *Jurnal Sisfokom (Sistem Informasi Dan Komputer)*, *12*(1), 105–116. https://doi.org/10.32736/sisfokom.v12i1.1614

Karne, N. (2022). Harnessing Sentiment Analytics: Insights into Customer Behavior and Decision-Making. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, 383–395. https://doi.org/10.32628/CSEIT228462

Kauffmann, E., Peral, J., Gil, D., Ferrández, A., Sellers, R., & Mora, H. (2019). Managing marketing decision-making with sentiment analysis: An evaluation of the main product features using text data mining. *Sustainability (Switzerland)*, *11*(15). https://doi.org/10.3390/su11154235

Khoirul, M., Hayati, U., & Nurdiawan, O. (2023). ANALISIS SENTIMEN APLIKASI BRIMO PADA ULASAN PENGGUNA DI GOOGLE PLAY MENGGUNAKAN ALGORITMA NAIVE BAYES. In *Jurnal Mahasiswa Teknik Informatika* (Vol. 7, Issue 1).

Kowsari, K., Meimandi, K. J., Heidarysafa, M., Mendu, S., Barnes, L., & Brown, D. (2019). Text classification algorithms: A survey. In *Information (Switzerland)* (Vol. 10, Issue 4). MDPI AG. https://doi.org/10.3390/info10040150

Lisnawati. (2024). *KEHADIRAN STARLINK DI INDONESIA: MANFAAT DAN DAMPAK*.

Manullang, O., Prianto, C., & Harani, N. H. (n.d.). *Analisis Sentimen Untuk Memprediksi Hasil Calon Pemilu Presiden Menggunakan Lexicon Based dan Random Forest*.

Mata, P. N. (n.d.). *SENTIMENT ANALYSIS-A LITERATURE REVIEW*.

Mata, P. N., & Martins, J. (n.d.). *Sentiment analysis: A literature review*. https://www.researchgate.net/publication/352553456

Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, *5*(4), 1093–1113. https://doi.org/10.1016/j.asej.2014.04.011

Mohan, N., Ferguson, A. E., Cech, H., Bose, R., Renatin, P. R., Marina, M. K., & Ott, J. (2024). A Multifaceted Look at Starlink Performance. *WWW 2024 - Proceedings of the ACM Web Conference*, 2723–2734. https://doi.org/10.1145/3589334.3645328

Muhammad, A. N. (2019). *Sentiment Analysis of Positive and Negative of YouTube Comments Using Naïve Bayes-Support Vector Machine (NBSVM) Classifier*.

Musfiroh, D., Khaira, U., Eko, P., Utomo, P., Suratno, T., Studi, P., Informasi, S., Sains, F., & Teknologi, D. (2021). *MALCOM: Indonesian Journal of Machine Learning and Computer Science Sentiment Analysis of Online Lectures in Indonesia from Twitter Dataset Using InSet Lexicon Analisis Sentimen terhadap Perkuliahan Daring di Indonesia dari Twitter Dataset Menggunakan InSet Lexicon*. *1*, 24–33.

Muslim, S. N. S., Nurdiyansyah, F., & Rahman, A. Y. (2024). PERBANDINGAN ALGORITMA NAIVE BAYES DAN KNN DALAM ANALISIS SENTIMEN ULASAN PENGGUNA APLIKASI CAPCUT. *Jurnal Informatika Dan Teknik Elektro Terapan*, *12*(3S1). https://doi.org/10.23960/jitet.v12i3s1.5156

Nisa, I. S. N., Rahmat Miyarno Saputro, Tegar Fatwa Nugroho, & Alfirna Rizqi Lahitani. (2024). Analisis Quality of Service (QoS) Menggunakan Standar Parameter Tiphon pada Jaringan Internet Berbasis Wi-Fi Kampus 1 Unjaya. *Teknomatika: Jurnal Informatika Dan Komputer*, *17*(1), 1–9. https://doi.org/10.30989/teknomatika.v17i1.1307

Paula, B., Fawzan, M., & Irsyad, H. (2024). Analisis Sentiment Masyarakat Terhadap penyebaran Starlink di Indonesia Menggunakan Algoritma Naive Bayes. *Journal Information & Computer JICOM*, *02*(2).

Putra, F. P. E., Mufidah, K., Ilhamsyah, R. M., Efendy, S. A., & Barokah, S. N. R. (2024). Tinjauan Performa RouterOS Mikrotik dalam Jaringan Internet: Analisis Kinerja dan Kelayakan. *Digital Transformation Technology*, *3*(2), 903–910. https://doi.org/10.47709/digitech.v3i2.3446

Rahman, M., & Munawaroh. (n.d.). *PENGARUH PENGGUNAAN MEDIA SOSIAL DAN DIFFERENTIATION PRODUK TERHADAP KEPUTUSAN PEMBELIAN PADA ERNI DIMSUM DI MEDAN JOHOR*.

Ramadhansyah, D., Asrofiq, A., & Yunefri, Y. (2024). Analisis Sentimen Ulasan penumpang maskapai penerbangan di Indonesia…ν ZONAsi. In *Jurnal Sistem Informasi* (Vol. 6, Issue 2).

Reza, M., Dores, A., Ambo, S. N., & Meilina, P. (2025). *PERBANDINGAN METODE MACHINE LEARNING UNTUK SENTIMEN ANALISIS REVIEW PENJUALAN PRODUK* (Vol. 15, Issue 2). https://jurnal.umj.ac.id/index.php/just-it/index

Syah, A., Nurdiyansyah, F., & Rahman, A. Y. (2024). ANALISIS SENTIMEN APLIKASI SHOPEE, TOKOPEDIA, LAZADA DAN BLIBLI MENGGUNAKAN LEKSIKON DAN RANDOM FOREST. *Jurnal Informatika Dan Teknik Elektro Terapan*, *12*(3S1). https://doi.org/10.23960/jitet.v12i3s1.5155

Syahril Dwi Prasetyo, Shofa Shofiah Hilabi, & Fitri Nurapriani. (2023). Analisis Sentimen Relokasi Ibukota Nusantara Menggunakan Algoritma Naïve Bayes dan KNN. *Jurnal KomtekInfo*, 1–7. https://doi.org/10.35134/komtekinfo.v10i1.330

Vargiu, E., & Urru, M. (2012). Exploiting web scraping in a collaborative filtering- based approach to web advertising. *Artificial Intelligence Research*, *2*(1). https://doi.org/10.5430/air.v2n1p44

Vens, C. (2013). Encyclopedia of Systems Biology. In *Encyclopedia of Systems Biology*. Springer New York. https://doi.org/10.1007/978-1-4419-9863-7

Widhiyasana, Y., Semiawan, T., Gibran, I., Mudzakir, A., & Noor, M. R. (2021). Penerapan Convolutional Long Short-Term Memory untuk Klasifikasi Teks Berita Bahasa Indonesia (Convolutional Long Short-Term Memory Implementation for Indonesian News Classification). In *Jurnal Nasional Teknik Elektro dan Teknologi Informasi* | (Vol. 10, Issue 4).

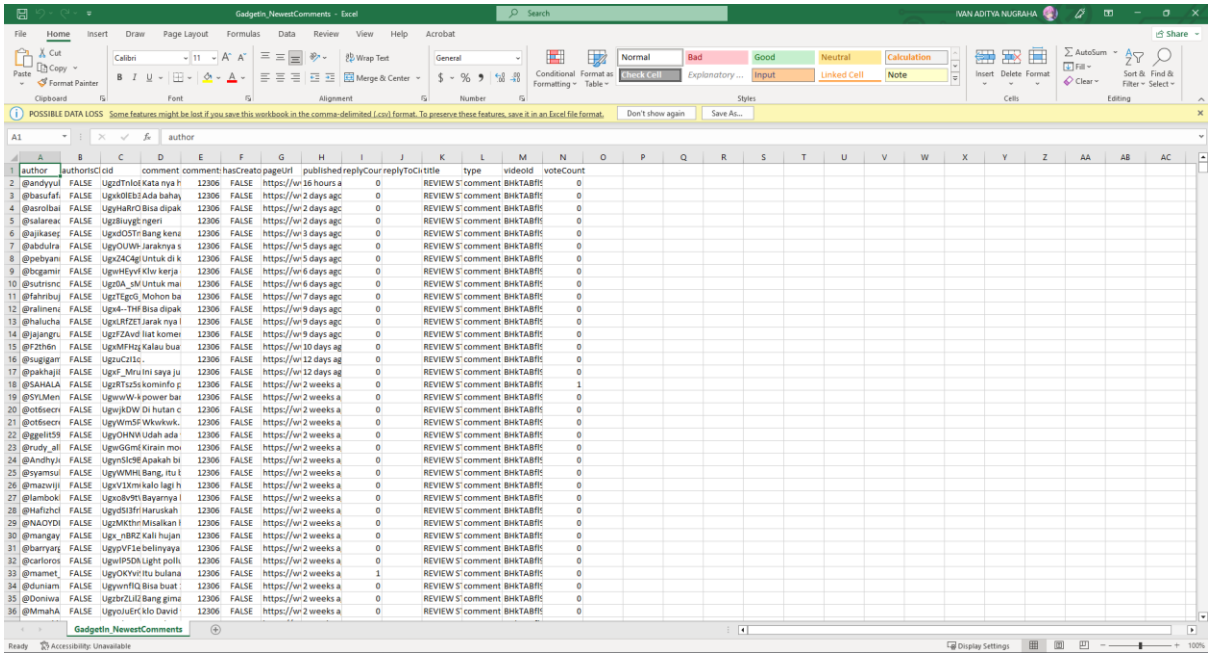
Yanti, C. P., Eva Agustini, N. W., Sri Rahayu Ginantra, N. L. W., & Putri Wulandari, D. A. (2023). Perbandingan Metode K-NN Dan Metode Random Forest Untuk Analisis Sentimen pada Tweet Isu Minyak Goreng di Indonesia. *JURNAL MEDIA INFORMATIKA BUDIDARMA*, *7*(2), 756. https://doi.org/10.30865/mib.v7i2.5900

Yolanda, N., Santi, I. H., Fanny, D., & Permadi, H. (2022). *Analisis Sentimen Popularitas Aplikasi Moodle Dan Edmodo Menggunakan Algoritma Support Vector Machine*. *3*(1), 48–59.

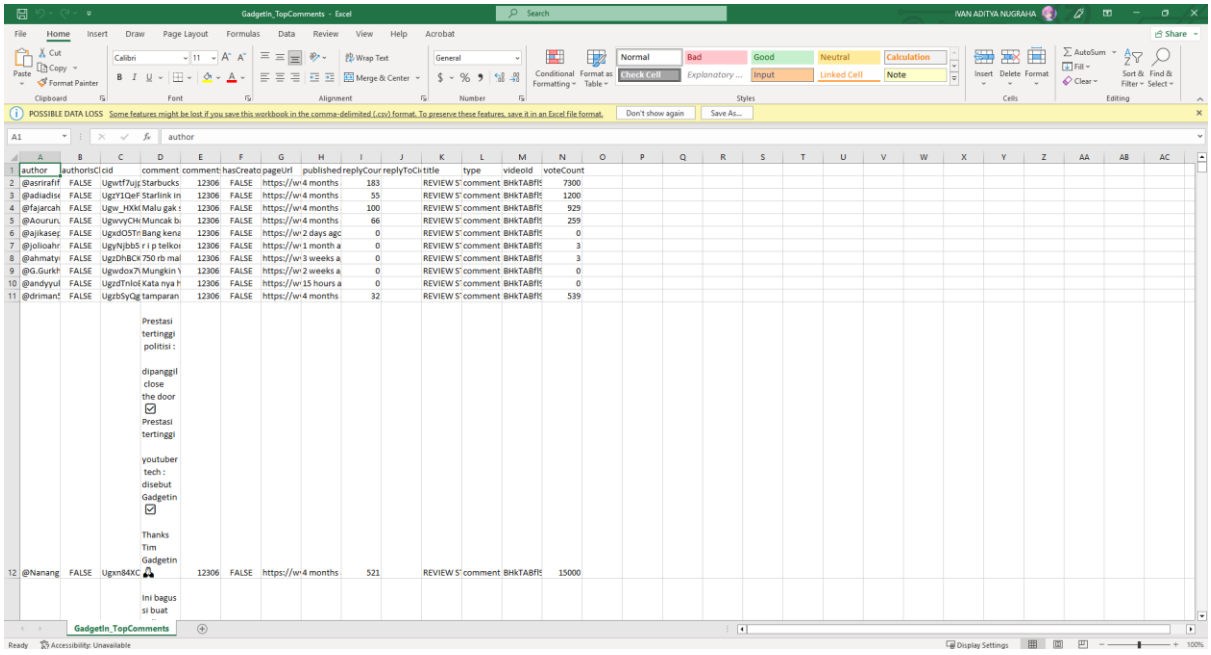
Zhi, H., Jiang, X., & Wang, J. (2024). Multicolour photometry of LEO mega-constellations Starlink and OneWeb. *Monthly Notices of the Royal Astronomical Society*, *530*(4), 5006–5015. https://doi.org/10.1093/mnras/stae693

LAMPIRAN

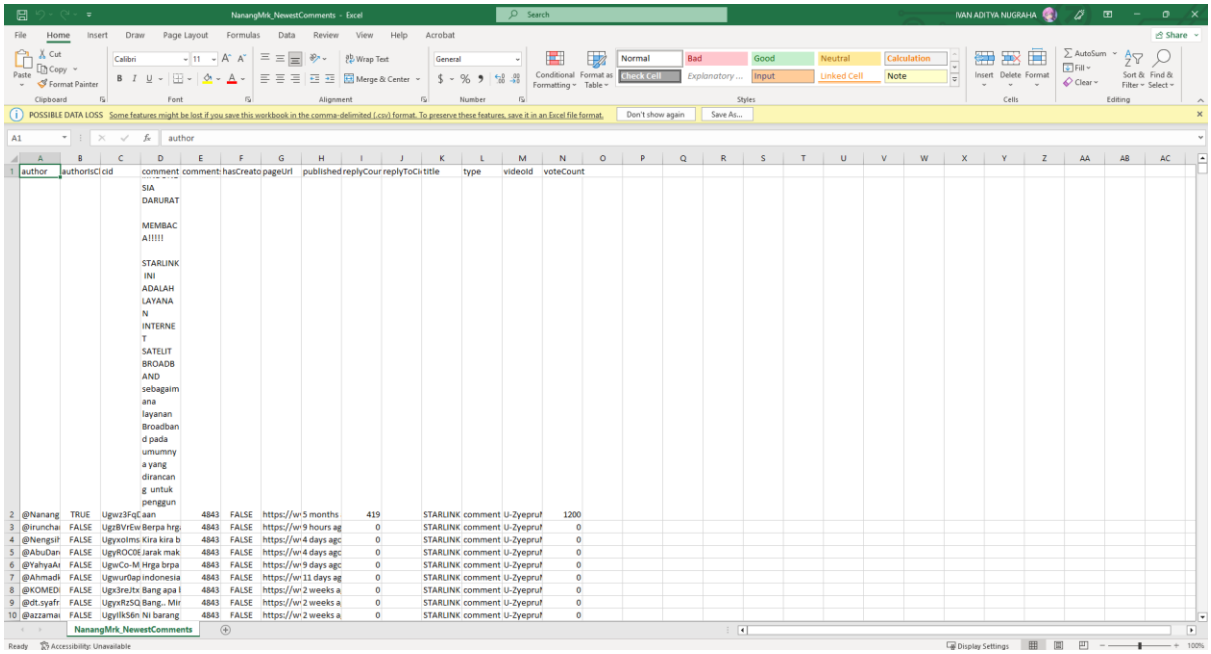
Dataset GadgetIn – *Newest Comments:*



Dataset GadgetIn – *Top Comments:*

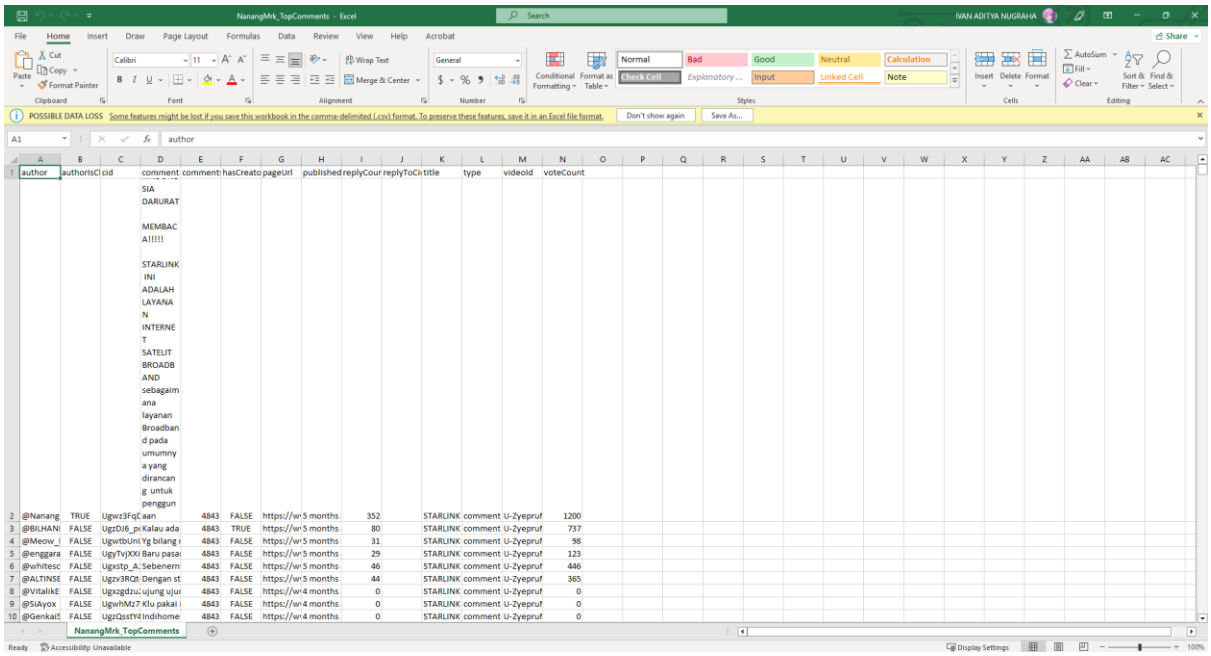


Dataset NanangMrk – *Newest Comments:*



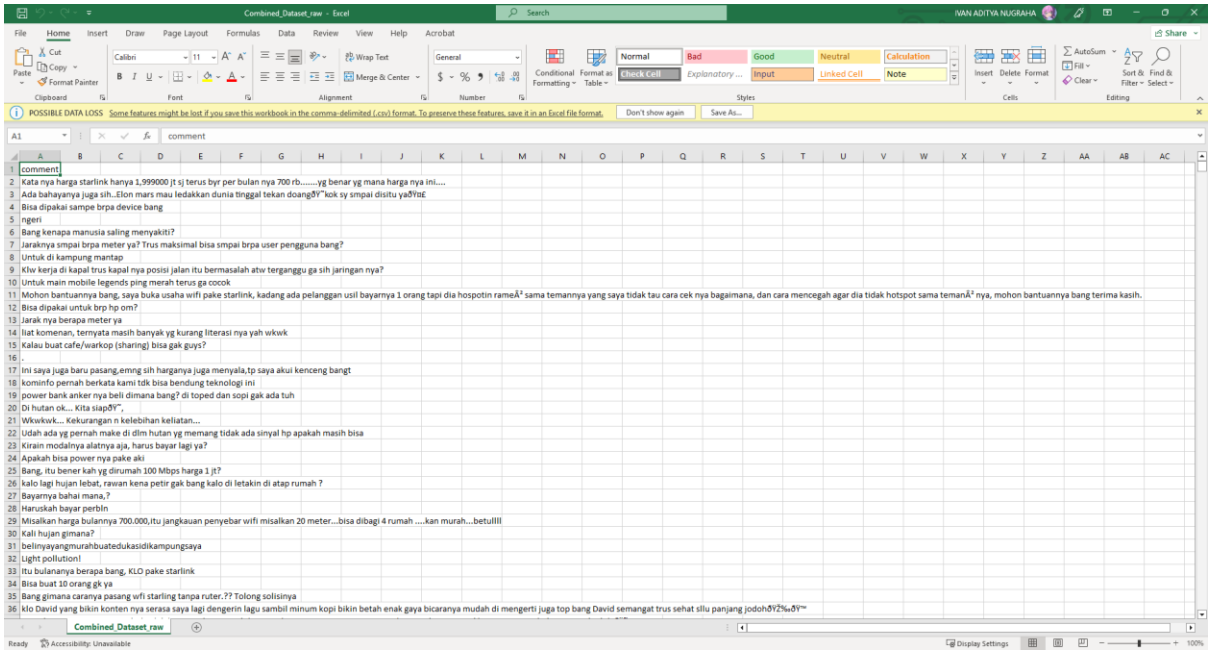
author	authorIsCl	cid	comment	comment:	hasCreato	pageUrl	published	replyCour	replyToCi	title	type	videoId	voteCount
			SIA DARURAT MEMBAC A!!!!! STARLINK INI ADALAH LAYANA N INTERNE T SATELIT BROADB AND sebagaim ana layanan Broadban d pada umumny a yang dirancan g untuk penggun										
@Nanang	TRUE	Ugwz3FqC	aan	4843	FALSE	https://w	5 months	419		STARLINK	comment	U-Zyepruf	1200
@irunchai	FALSE	UgzBVrEw	Berpa hrg	4843	FALSE	https://w	9 hours ag	0		STARLINK	comment	U-Zyepruf	0
@Nengsih	FALSE	UgyxoIms	Kira kira b	4843	FALSE	https://w	4 days agc	0		STARLINK	comment	U-Zyepruf	0
@AbuDan	FALSE	UgyROC0E	Jarak mak	4843	FALSE	https://w	4 days agc	0		STARLINK	comment	U-Zyepruf	0
@YahyaAs	FALSE	UgwCo-M	Hrga brpa	4843	FALSE	https://w	9 days agc	0		STARLINK	comment	U-Zyepruf	0
@Ahmadk	FALSE	Ugwur0ap	indonesia	4843	FALSE	https://w	11 days ag	0		STARLINK	comment	U-Zyepruf	0
@KOMEDI	FALSE	Ugx3reJtx	Bang apa l	4843	FALSE	https://w	2 weeks a	0		STARLINK	comment	U-Zyepruf	0
@dt.syafr	FALSE	UgyxRzSQ	Bang.. Mir	4843	FALSE	https://w	2 weeks a	0		STARLINK	comment	U-Zyepruf	0
@azzamai	FALSE	Ugyllk56n	Ni barang	4843	FALSE	https://w	2 weeks a	0		STARLINK	comment	U-Zyepruf	0

Dataset NanangMrk – *Top Comments:*



author	authorIsCl	cid	comment	comment:	hasCreato	pageUrl	published	replyCour	replyToCi	title	type	videoId	voteCount
			SIA DARURAT MEMBAC A!!!!! STARLINK INI ADALAH LAYANA N INTERNE T SATELIT BROADB AND sebagaim ana layanan Broadban d pada umumny a yang dirancan g untuk penggun										
@Nanang	TRUE	Ugwz3FqC	aan	4843	FALSE	https://w	5 months	352		STARLINK	comment	U-Zyepruf	1200
@BILHANI	FALSE	UgzDJ6_pi	Kalau ada	4843	TRUE	https://w	5 months	80		STARLINK	comment	U-Zyepruf	737
@Meow_I	FALSE	UgwtbUnl	Yg bilang r	4843	FALSE	https://w	5 months	31		STARLINK	comment	U-Zyepruf	98
@enggara	FALSE	UgyTvjXXi	Baru pasai	4843	FALSE	https://w	5 months	29		STARLINK	comment	U-Zyepruf	123
@whitesc	FALSE	Ugxstp_A	Sebenern	4843	FALSE	https://w	5 months	46		STARLINK	comment	U-Zyepruf	446
@ALTINSE	FALSE	Ugzv3RQt	Dengan st	4843	FALSE	https://w	5 months	44		STARLINK	comment	U-Zyepruf	365
@VitaliikE	FALSE	Ugxzgdzu	ujung ujur	4843	FALSE	https://w	4 months	0		STARLINK	comment	U-Zyepruf	0
@SiAyox	FALSE	UgwhMz7	Klu pakai i	4843	FALSE	https://w	4 months	0		STARLINK	comment	U-Zyepruf	0
@Genkai5	FALSE	UgzQsstY	Indihome	4843	FALSE	https://w	4 months	0		STARLINK	comment	U-Zyepruf	0

Dataset gabungan:



comment
Kata nya harga starlink hanya 1,999000 jt sj terus byr per bulan nya 700 rb.....yg benar yg mana harga nya ini....
Ada bahayanya juga sih..Elon mars mau ledakkan dunia tinggal tekan doangð٘ kok sy smpai disitu yað٘
Bisa dipakai sampe brpa device bang
ngeri
Bang kenapa manusia saling menyakiti?
Jaraknya smpai brpa meter ya? Trus maksimal bisa smpai brpa user pengguna bang?
Untuk di kampung mantap
Klw kerja di kapal trus kapal nya posisi jalan itu bermasalah atw terganggu ga sih jaringan nya?
Untuk main mobile legends ping merah terus ga cocok
Mohon bantuannya bang, saya buka usaha wifi pake starlink, kadang ada pelanggan usil bayarnya 1 orang tapi dia hospotin rameÃ² sama temannya yang saya tidak tau cara cek nya bagaimana, dan cara mencegah agar dia tidak hotspot sama temanÃ² nya, mohon bantuannya bang terima kasih.
Bisa dipakai untuk brp hp om?
Jarak nya berapa meter ya
liat komenan, ternyata masih banyak yg kurang literasi nya yah wkwk
Kalau buat cafe/warkop (sharing) bisa gak guys?
.
Ini saya juga baru pasang,emng sih harganya juga menyala,tp saya akui kenceng bangt
kominfo pernah berkata kami tdk bisa bendung teknologi ini
power bank anker nya beli dimana bang? di toped dan sopi gak ada tuh
Di hutan ok... Kita siapð٘
Wkwkwk... Kekurangan n kelebihan keliatan...
Udah ada yg pernah make di dlm hutan yg memang tidak ada sinyal hp apakah masih bisa
Kirain modalnya alatnya aja, harus bayar lagi ya?
Apakah bisa power nya pake aki
Bang, itu bener kah yg dirumah 100 Mbps harga 1 jt?
kalo lagi hujan lebat, rawan kena petir gak bang kalo di letakin di atap rumah ?
Bayarnya bahai mana,?
Haruskah bayar perbln
Misalkan harga bulannya 700.000,itu jangkauan penyebar wifi misalkan 20 meter...bisa dibagi 4 rumah ....kan murah...betullll
Kali hujan gimana?
belinyayangmurahbuatedukasidikampungsaya
Light pollution!
Itu bulananya berapa bang, KLO pake starlink
Bisa buat 10 orang gk ya
Bang gimana caranya pasang wfi starling tanpa ruter.?? Tolong solisinya
klo David yang bikin konten nya serasa saya lagi dengerin lagu sambil minum kopi bikin betah enak gaya bicaranya mudah di mengerti juga top bang David semangat trus sehat sllu panjang jodohð٘