INNOVATIVE: Journal Of Social Science Research
Volume 5 Nomor 3 Tahun 2025 Page 612-626
E-ISSN 2807-4238 and P-ISSN 2807-4246
Website: https://j-innovative.org/index.php/Innovative

Analisis Sentimen Video Review Starlink Indonesia Menggunakan Pendekatan Lexicon Dictionary Bahasa Indonesia dan Algoritma Random Forest

Ivan Aditya Nugraha[1✉], Irving Vitra Paputungan[2]
Universitas Islam Indonesia
Email: 20523080@students.uii.ac.id[1✉]

Abstrak

Starlink dari SpaceX menawarkan solusi internet satelit bagi wilayah tertinggal, terdepan, dan terluar (3T) di Indonesia yang masih minim akses digital. Penelitian ini bertujuan menganalisis tanggapan publik terhadap layanan Starlink melalui 9.372 komentar YouTube dari kanal GadgetIn dan NanangMrk. Data dikumpulkan menggunakan dan diproses melalui tahapan pembersihan teks, normalisasi, tokenisasi, penghapusan stopword, serta stemming. Sentimen pengguna diklasifikasikan ke dalam kategori positif, negatif, dan netral menggunakan pendekatan lexicon Bahasa Indonesia. Selanjutnya, model klasifikasi dibangun menggunakan algoritma Random Forest dengan tiga skenario rasio data latih dan uji, yaitu 80:20, 70:30, dan 60:40. Hasil evaluasi menunjukkan performa terbaik pada rasio 80:20, dengan akurasi 77,4%, precision 0,78, recall 0,77, dan F1-score 0,77 berdasarkan pengujian terhadap data uji.

Kata Kunci: *Analisis Sentimen, Lexicon Dictionary Bahasa Indonesia, Random Forest*

Abstract

Starlink by SpaceX offers a satellite internet solution for underdeveloped, frontier, and outermost (3T) regions in Indonesia that still have limited digital access. This study aims to analyze public responses to the Starlink service through 9,372 YouTube comments from the GadgetIn and NanangMrk channels. The data was collected and processed through text cleaning, normalization, tokenization, stopword removal, and stemming stages. User sentiment was classified into positive, negative, and neutral categories using an Indonesian lexicon approach. Subsequently, a classification model was built using the Random Forest algorithm with three training and testing data ratio scenarios: 80:20, 70:30, and 60:40. The evaluation results show the best performance at the 80:20 ratio, with an accuracy of 77.4%, precision of 0.78, recall of 0.77, and an F1-score of 0.77 based on testing against the test data.Keyword: Indonesian lexicon dictionary, random forest, sentiment analysis.

Keywords: *Indonesian Lexicon Dictionary, Sentiment Analysis, Random Forest*

PENDAHULUAN

Dalam era digital saat ini, akses terhadap internet telah menjadi kebutuhan mendasar bagi masyarakat di seluruh dunia, termasuk Indonesia. Namun, tantangan geografis dan infrastruktur yang terbatas menyebabkan kesenjangan digital yang signifikan, terutama di wilayah 3T (terdepan, terluar, dan tertinggal). Menurut (Rahman & Munawaroh, t.t.) berdasarkan hasil survei Asosiasi Penyelenggara Jasa Internet Indonesia (APJII), jumlah pengguna internet di Indonesia mencapai 215,63 juta jiwa pada periode tahun 2022–2023, menunjukkan bahwa mayoritas penduduk telah mengandalkan internet dalam aktivitas sehari – hari.

Kehadiran Starlink, layanan internet berbasis satelit orbit rendah *(Low Earth Orbit/LEO)* yang dikembangkan oleh SpaceX, menawarkan solusi potensial untuk mengatasi permasalahan ini dengan menyediakan konektivitas internet berkecepatan tinggi dan latensi rendah di daerah-daerah terpencil (Lisnawati, 2024). Starlink memiliki tujuan, yaitu menyediakan internet global melalui armada satelit yang terbang sekitar 500 km di atas permukaan bumi (Mohan dkk., 2024). Berbeda dari jaringan internet konvensional yang mengandalkan infrastruktur berbasis darat, Starlink memanfaatkan ribuan satelit kecil yang saling terhubung untuk menciptakan jaringan global. Pendekatan ini memungkinkan penyediaan layanan internet hingga ke wilayah – wilayah yang sebelumnya sulit dijangkau, seperti daerah terpencil atau pedesaan.

Pada tanggal 19 Mei 2024, Elon Musk secara resmi menginformasikan peluncuran Starlink di Indonesia melalui unggahan di akun pribadinya di platform X (sebelumnya Twitter). Dalam pernyataannya yang menunjukkan penghargaan atas peluncuran ini,

momen tersebut menjadi tonggak dimulainya era baru konektivitas internet di Indonesia, khususnya untuk daerah – daerah yang selama ini sulit dijangkau oleh infrastruktur konvensional.

Di Indonesia, antusiasme terhadap layanan Starlink terlihat dari berbagai ulasan dan diskusi yang muncul di platform digital, khususnya YouTube. Salah satu konten yang menonjol adalah video dari kanal NanangMrk berjudul "STARLINK INDONESIA REVIEW, INTERNET DARI LANGIT TANPA SINYAL DIMANAPUN ANDA BERADA (TERLENGKAP)", yang diunggah pada 11 Mei 2024 yang dimana merupakan video *review* Starlink di Indonesia yang pertama kali. 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.

Selain NanangMrk, kanal YouTube teknologi populer lainnya, GadgetIn, juga membahas tentang layanan Starlink melalui video berjudul "Review Starlink Indonesia – Ini Revolusioner". Video ini menjadi video *review* Starlink di Indonesia dengan jumlah penonton terbanyak. Yang dimana video ini telah ditonton sebanyak lebih dari 3,6 juta kali serta mendapat puluhan ribu like dari penonton. Dalam video berdurasi 24 menit ini, pemilik kanal GadgetIn, David Brendi, 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 yang telah ada.

Respon publik terhadap dua video ulasan tersebut sangat beragam, mulai dari komentar yang bernada antusias, kritik terhadap harga dan ketersediaan layanan, hingga pertanyaan teknis seputar penggunaan Starlink. Untuk memahami pola opini tersebut secara sistematis, pendekatan analisis sentimen dapat digunakan. Analisis sentimen adalah aktivitas penyelidikan terhadap opini yang diungkapkan oleh masyarakat mengenai suatu peristiwa atau fenomena. Metode ini juga dikenal sebagai *opinion mining*, bagian dari studi komputasi dalam *Natural Language Processing* (NLP), yang bertujuan mengekstrak nilai sentimen dari suatu teks. Sumber data dapat berasal dari berbagai kanal digital seperti ulasan produk, berita, hingga komentar media sosial. Nilai sentimen tersebut dapat dikategorikan ke dalam tiga kelas: positif, netral, atau negatif (Ahmad & Gata, 2022).

Selain pendekatan analisis sentimen, proses memahami data dalam jumlah besar dari komentar YouTube juga melibatkan tahapan klasifikasi teks. Klasifikasi teks adalah proses

mengelompokkan teks ke dalam beberapa kategori tertentu, seperti sentimen positif, netral, atau negatif. Proses ini mencakup ekstraksi fitur, reduksi dimensi, pemilihan algoritma, dan evaluasi model (Widhiyasana dkk., 2021). Dalam konteks ini, algoritma klasifikasi seperti *random forest* banyak digunakan karena keandalannya dalam menghasilkan akurasi tinggi serta kemampuannya dalam menangani data yang tidak seimbang atau mengandung outlier (Manullang dkk., t.t.).

METODE PENELITIAN

Penelitian ini bertujuan untuk menganalisis dan memprediksi sentimen berdasarkan data komentar *netizen* di media sosial YouTube yang berkaitan dengan ulasan layanan internet Starlink di Indonesia. Data yang digunakan merupakan data sekunder yang diperoleh melalui proses scraping menggunakan *platform* Apify.com, dengan fokus pada dua video ulasan Starlink dari kanal GadgetIn dan NanangMrk. Penelitian ini dilaksanakan melalui lima tahapan utama. Tahap awal adalah pengumpulan data dari kedua video tersebut. Selanjutnya, dilakukan tahap *preprocessing* yang mencakup pembersihan data, penghapusan duplikasi, hingga proses *stemming* untuk mempersiapkan data agar sesuai dengan format analisis. Setelah itu, data diberi label sentimen dan diproses lebih lanjut untuk divisualisasikan menggunakan *wordcloud* berdasarkan kategori sentimen. Pada tahap pelatihan, data dibagi ke dalam beberapa rasio (60:40, 70:30, dan 80:20) untuk menguji performa model. Algoritma *random forest* digunakan dalam proses pelatihan ini. Terakhir, dilakukan evaluasi model berdasarkan metrik akurasi, *presisi*, *recall*, dan *f1-score* untuk menilai kinerja model prediksi sentimen.

Data Collection

Data diperoleh dari komentar pengguna pada dua video YouTube yang membahas layanan Starlink di Indonesia, masing – masing dari kanal GadgetIn dan NanangMrk. Video dari GadgetIn dipilih karena memiliki jumlah penonton terbanyak untuk topik Starlink, sementara video dari NanangMrk dipilih karena menjadi video ulasan pertama tentang Starlink di Indonesia. Dari masing – masing video dikumpulkan dua kategori komentar, yaitu *top comments* dan *newest comments*, dengan batas maksimal 2.500 komentar per kategori karena keterbatasan akses pengambilan data.

Berikut adalah dataset untuk video "Review Starlink Indonesia – Ini Revolusioner" dari kanal GadgetIn:

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

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 1. Dataset komentar kanal GadgetIn

Berikut adalah dataset untuk video "STARLINK INDONESIA REVIEW, INTERNET DARI LANGIT TANPA SINYAL DIMANAPUN ANDA BERADA (TERLENGKAP)" dari kanal NanangMrk:

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

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 2. Dataset komentar kanal NanangMrk

Empat dataset yang terkumpul kemudian digabung menjadi satu dataset gabungan yang berisi total 9.372 komentar. Namun, dalam proses penggabungan ditemukan adanya komentar ganda karena beberapa komentar muncul di dua kategori sekaligus. Oleh karena

itu, sebelum dianalisis lebih lanjut, data terlebih dahulu melalui tahap praproses untuk menghapus duplikasi dan membersihkan data yang tidak relevan.

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. Dataset gabungan

Preprocessing

Preprocessing data adalah langkah awal dalam pengolahan data yang bertujuan untuk menyiapkan data mentah agar siap digunakan dalam proses pemodelan. Tahapan ini umumnya mencakup pembersihan dan transformasi data dasar guna memastikan data sesuai dengan format atau kebutuhan input dari algoritma yang akan diterapkan.

1. Data Cleaning

Data komentar terlebih dahulu diperiksa untuk memastikan tidak terdapat duplikasi yang dapat memengaruhi hasil analisis. Setelah itu, dilakukan proses pembersihan data yang mencakup penghapusan karakter tidak relevan seperti mention (@), hashtag (#), tautan URL, angka, tanda baca, serta karakter – karakter khusus lainnya yang tidak memberikan kontribusi makna terhadap analisis sentimen. Lalu, semua kata dalam teks dirubah menjadi huruf kecil.

	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 4. Dataset setelah dibersihkan

2. Normalization

Normalisasi merupakan tahap awal dalam pemrosesan teks yang bertujuan untuk menyeragamkan teks mentah agar lebih konsisten dan mudah dianalisis. Proses ini mencakup pengubahan kata-kata ke dalam bentuk baku. Dalam bahasa Indonesia, normalisasi juga mencakup konversi kata tidak baku, slang, atau singkatan menjadi bentuk yang sesuai dengan kaidah bahasa. Adanya singkatan, kesalahan penulisan (typo), dan bahasa gaul dalam dataset dapat memengaruhi hasil analisis, sehingga normalisasi berperan penting dalam meningkatkan akurasi dan kualitas analisis data.

	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

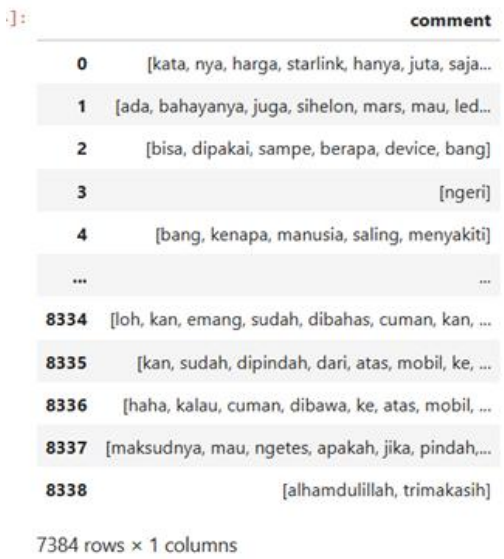
7384 rows × 1 columns

Gambar 5. Dataset setelah melalui 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.*

3. Tokenizing

Tokenisasi, atau yang dikenal sebagai tokenization, adalah proses memisahkan teks menjadi unit-unit kecil berupa kata – kata individu yang disebut "token." Langkah ini sangat penting dalam pemrosesan teks karena memungkinkan algoritma untuk mengenali dan memahami komponen dasar dari teks yang akan dianalisis.



	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 6. Dataset setelah melalui tokenisasi

4. *Stopword Removal*

Stopword removal adalah proses menghilangkan kata – kata umum yang sering muncul dalam teks, namun biasanya tidak memiliki makna dalam analisis, seperti "dan," "di," "yang," atau "untuk" dalam bahasa Indonesia. Kata – kata semacam ini dikenal sebagai stopwords. Karena stopwords cenderung tidak membawa informasi penting dalam analisis teks atau sentimen, penghapusannya membantu mengurangi gangguan (noise) dalam data dan mempercepat proses analisis. Dengan menghapus kata-kata tersebut, model dapat lebih fokus pada kata-kata yang relevan dan bermakna, sehingga dapat meningkatkan akurasi dalam pemrosesan dan interpretasi data.

	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 7. Dataset setelah melalui stopword removal

5. Stemming

Terakhir, dilakukan proses stemming untuk mengembalikan setiap kata ke bentuk dasarnya. Proses ini berguna agar model dapat mengenali kata yang memiliki akar yang sama sebagai satu entitas, misalnya “berlari”, “lari”, dan “pelari” dikembalikan ke kata dasar “lari”.

	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 8. Dataset setelah melalui 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.

Selanjutnya dataset yang telah dibersihkan disimpan, lalu lanjut ke dalam tahapan processing.

Processing

Data yang telah dikumpulkan melalui proses scraping memiliki format yang tidak terstruktur dan sulit dibaca oleh mesin, karena mencakup seluruh variabel yang tersedia (Paula dkk., 2024). Oleh karena itu, diperlukan tahapan preprocessing untuk membersihkan dan menyesuaikan data sebelum dianalisis. Terdapat dua tahapan dalam processing, yaitu pelabelan sentimen dan visualisasi.

1. Pelabelan Sentimen

Pelabelan sentimen merupakan proses mengklasifikasikan emosi atau opini yang terkandung dalam suatu teks ke dalam kategori tertentu. Tujuannya adalah untuk mengenali dan menentukan sikap penulis terhadap suatu topik berdasarkan kata-kata yang digunakan.

Pelabelan sentimen akan diklasifikasikan ke dalam tiga kategori, yaitu positif, negatif, dan netral. Untuk mempermudah proses pelabelan, digunakan kamus lexicon berbahasa Indonesia sebagai acuan. Kamus ini berisi daftar kata yang telah diberi nilai sentimen dan berfungsi sebagai referensi dalam menentukan sentimen suatu teks atau kalimat berdasarkan kata-kata yang terkandung di dalamnya.

	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 9. Kalimat dengan nilai sentimenya

Apabila data dibagi berdasarkan kategori sentimennya, maka terdapat 2.627 data dengan sentimen positif, 3.296 data dengan sentimen negatif, dan 1.461 data yang termasuk dalam kategori netral.

Tabel 1. Total data per nilai sentimen

Sentimen	Total Data
Positif	2.627
Negatif	3.296
Netral	1.461

2. Visualisasi

Pada penelitian ini, visualisasi dilakukan menggunakan *wordcloud*, yaitu representasi data berupa sekumpulan kata yang ditata secara acak dengan ukuran berbeda – beda. Ukuran setiap yangkata menunjukkan frekuensi kemunculannya dalam dataset. Semakin sering kata muncul, maka semakin besar ukurannya (Iqbal & Yusuf, t.t.)

Berikut adalah visualisasi kata yang bernilai sentimen positif, negatif, dan netral dalam bentuk *wordcloud*

Gambar 10. Wordcloud masing - masing nilai sentimen

Pada wordcloud sentimen positif, kata-kata seperti "Starlink", "paket", "rumah", dan "mantap" sering muncul dan ditampilkan dengan ukuran lebih besar. Ini menunjukkan bahwa banyak pengguna memberi tanggapan positif terhadap layanan Starlink, terutama dalam hal koneksi yang stabil dan kemudahan penggunaan. Kata seperti "mau", "coba", dan "pakai" juga menunjukkan minat pengguna untuk mencoba layanan ini.

Sementara itu, wordcloud sentimen negatif memperlihatkan kata-kata seperti "Starlink", "sinyal", "Indonesia", "bayar", dan "pakai" yang sering muncul dalam komentar bernada negatif. Hal ini menunjukkan adanya keluhan dari pengguna terkait masalah sinyal, harga, dan penggunaan layanan. Kata seperti "bukan", "jangan", dan "mahal" mencerminkan rasa kecewa atau ketidakpuasan.

Untuk wordcloud sentimen netral, kata-kata yang muncul antara lain "starlink", "internet", "pakai", "berapa", "wifi", dan "review". Ini menunjukkan bahwa banyak komentar bersifat informatif atau berupa pertanyaan, tanpa menunjukkan pendapat yang kuat. Pengguna lebih fokus pada informasi teknis, harga, dan perbandingan dengan layanan lain.

Training Data

Pembagian data menggunakan teknik *hold-out validation*, yaitu membagi data menjadi dua set utama, yaitu data latih (training data) dan data uji (testing data). Teknik ini sederhana, namun terbukti efektif dan sering digunakan (Putri dkk., 2024). Untuk menguji

stabilitas dan performa algoritma *random forest* dalam klasifikasi sentimen, dilakukan eksperimen dengan tiga variasi pembagian data: 60:40, 70:30, dan 80:20. Model dilatih menggunakan data latih dan dievaluasi pada data uji, dengan pemisahan dilakukan secara acak menggunakan parameter random_state = 42 untuk memastikan hasil yang konsisten. Proses pelatihan dilakukan dengan pipeline. Model dilatih menggunakan data latih, lalu diuji dengan data uji untuk menghasilkan nilai akurasi, *precision*, *recall*, dan *F1-score* pada masing-masing kelas sentimen.

HASIL DAN PEMBAHASAN

Evaluasi Model

Setelah data dilatih, kemudian masuk ke tahapan evaluasi model. Evaluasi model ini melibatkan matriks yang Bernama *confusion matrix* untuk melihat prediksi mana yang benar dan yang salah. *Confusion matrix* sendiri adalah matriks yang digunakan untuk mengevaluasi performa model klasifikasi dalam *machine learning.* Matriks ini membandingkan prediksi model dengan nilai actual *(ground truth)* dan membagi hasilnya menjadi empat kategori, yaitu *True Positive (TP), True Negative (TN), False Positive (FP),* dan *False Negative (FN).* Berikut adalah penjelasan masing – masing:

1. *True Positive (TP),* jumlah sampel yang diprediksi positif oleh model dan memang benar positf sesuai data actual,
2. *True Negative (TN),* jumlah sampel yang diprediksi negative oleh model dan memang benar – benar negative 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 negative oleh model, tetapi sebenarnya positif.

Nilai – nilai akurasi, *precision, recall,* dan *F1-score* dapat diperoleh dari *confusion matrix* untuk menilai kinerja model yang telah dibangun. Setiap model yang telah melalui proses pelatihan dievaluasi menggunakan keempat matrik tersebut. Akurasi berfungsi untuk menunjukkan seberapa sering model membuat prediksi yang tepat secara keseluruhan, sedangkan *precision* dan *recall* digunakan untuk menilai kinerja model dalam mengklasifikasikan masing – masing kategori sentimen (positif, negatif, dan netral). Adapun *F1-score* merupakan rata – rata harmonik dari *precision* dan *recall*, yang memberikan gambaran menyeluruh dan seimbang mengenai performa model.

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

a. *Accuracy*

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

b. *Precision*

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

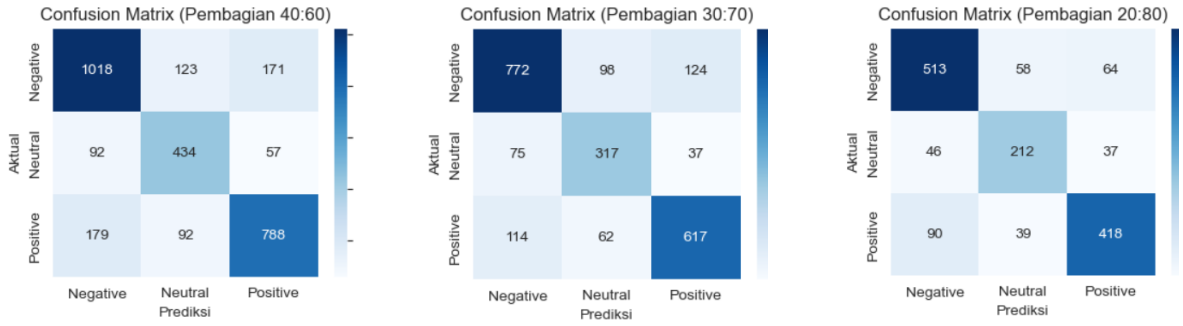
c. *Recall*

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

d. *F1-score*

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

Berikut adalah hasil dari tahapan *training data* dalam bentuk *confusion matrix:*



Gambar 11. Confusion matrix dari ketiga raiso

Tabel berikut merangkum performa model berdasarkan rasio pembagian data:

Tabel 2. Rangkuman performa model

Rasio Data	*Accuracy*	*Precision*	*Recall*	*F1-Score*
60:40	0.758	0.760	0.758	0.759
70:30	0.769	0.772	0.769	0.770
80:20	0.773	0.775	0.773	0.774

Hasil evaluasi performa model menunjukkan bahwa algoritma Random Forest yang digunakan dalam analisis sentimen ini memberikan performa terbaik pada rasio pembagian data dan data uji sebesar 80:20. Pada rasio tersebut, model mencapai akurasi sebesar

77,4%, dengan nilai *precision* rata – rata 0,78, *recall* 0,77, dan *F1-score* sebesar 0,77. Nilai – nilai ini diperoleh berdasarkan penghitungan metrik evaluasi dari *confusion matrix* dan menunjukkan bahwa model mampu melakukan klasifikasi sentimen dengan cukup baik pada ketiga kategori, yakni positif, negatif, dan netral.

Jika dibandingkan dengan rasio pembagian data 70:30 dan 60:40, rasio 80:20 secara konsisten menghasilkan nilai tertinggi untuk seluruh metrik evaluasi. Hal ini mengindikasikan bahwa model *random forest* dengan 200 *estimator* yang digunakan mampu mengenali pola sentimen secara lebih efektif ketika proporsi data latih lebih besar, dikarenakan variasi data latih yang lebih kaya memberikan pengukuran performa yang lebih representatif.

SIMPULAN

Penelitian ini menunjukkan bahwa algoritma Random Forest efektif untuk analisis sentimen komentar YouTube berbahasa Indonesia, dengan performa terbaik pada rasio data latih dan uji 80:20. Temuan ini memajukan pemahaman tentang penerapan *machine learning* berbasis *lexicon* dalam konteks bahasa lokal, khususnya untuk klasifikasi tiga kelas sentiment dengan hasil metrik evaluasi yang stabil dan cukup tinggi. Penelitian selanjutnya disarankan mengeksplorasi tuning hiperparameter lanjutan dan membandingkan model ini dengan pendekatan berbasis deep learning untuk meningkatkan akurasi klasifikasi.

DAFTAR PUSTAKA

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

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

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

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

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

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).

Putri, S. M., Novita, R., & Afdal, M. (2024). Perbandingan Algoritma Linear Regression, Support Vector Regression, dan Artificial Neural Network untuk Prediksi Data Obat. Technology and Science (BITS), 6(1). https://doi.org/10.47065/bits.v6i1.5184

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

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). Dalam Jurnal Nasional Teknik Elektro dan Teknologi Informasi | (Vol. 10, Nomor 4).