ANALISIS SENTIMEN KOMENTAR PENONTON YOUTUBE TERHADAP TOPIK KORUPSI PADA KANAL EDUKASI MALAKA PROJECT MENGGUNAKAN ALGORITMA *RANDOM FOREST*

SKRIPSI

DISUSUN OLEH

MUHAMMAD DIMAS ALFATHIR

NIM.220607110026

PROGRAM STUDI PERPUSTAKAAN DAN SAINS INFORMASI

FAKULTAS SAINS DAN TEKNOLOGI

UNIVERSITAS ISLAM NEGERI MAULANA MALIK IBRAHIM MALANG

2026

HALAMAN JUDUL

ANALISIS SENTIMEN KOMENTAR PENONTON YOUTUBE TERHADAP TOPIK KORUPSI PADA KANAL EDUKASI MALAKA PROJECT MENGGUNAKAN ALGORITMA *RANDOM FOREST*

SKRIPSI

Disusun Oleh:

MUHAMMAD DIMAS ALFATHIR

NIM.220607110026

Diajukan Kepada:
Fakultas Sains dan Teknologi
Universitas Islam Negeri Maulana Malik Ibrahim Malang
untuk Memenuhi Salah Satu Persyaratan dalam
Memperoleh Gelar Sarjana Sains Informasi (S.S.I)

PROGRAM STUDI PERPUSTAKAAN DAN SAINS INFORMASI

FAKULTAS SAINS DAN TEKNOLOGI

UNIVERSITAS ISLAM NEGERI MAULANA MALIK IBRAHIM

MALANG

2026

HALAMAN PERSETUJUAN

ANALISIS SENTIMEN KOMENTAR PENONTON YOUTUBE TERHADAP TOPIK KORUPSI PADA KANAL EDUKASI MALAKA PROJECT MENGGUNAKAN ALGORITMA *RANDOM FOREST*

SKRIPSI

Oleh:

MUHAMMAD DIMAS ALFATHIR
NIM. 220607110026

Telah Diperiksa dan Disetujui untuk Diuji:
Tanggal: 22 April 2026

Pembimbing I,	Pembimbing II,
Fakhris Khusnu Reza Mahfud, M.Kom	**Yulianto, M.Pd.I**
NIP. 199005062019031007	**NIP. 198707122019031005**

Mengetahui,
Ketua Program Studi Perpustakaan dan Sains Informasi
Fakultas Sains dan Teknologi
Universitas Islam Negeri Maulana Malik Ibrahim Malang

Dra. Siti Mudawamah, M.IP
NIP. 196002232018012001

HALAMAN PENGESAHAN

ANALISIS SENTIMEN KOMENTAR PENONTON YOUTUBE TERHADAP TOPIK KORUPSI PADA KANAL EDUKASI MALAKA PROJECT MENGGUNAKAN ALGORITMA *RANDOM FOREST*

SKRIPSI

Oleh:

MUHAMMAD DIMAS ALFATHIR

NIM.220607110026

Telah dipertahankan di depan Dewan Penguji Skripsi dan dinyatakan diterima sebagai salah satu persyaratan untuk memperoleh Gelar Sarjana Sains Informasi (S.S.I) pada tanggal 22 April 2026

Susunan Dewan Penguji		**Tanda Tangan**
Ketua Penguji	: **Firma Sahrul Bahtiar, M. Eng.** **NIP. 198502012019031009**	
Anggota Penguji I	: **Annisa Fajriyah, M.A** **NIP. 198801122020122002**	
Anggota Penguji II	: **Fakhris Khusnu Reza Mahfud, M.Kom** **NIP. 199005062019031007**	
Anngota Penguji III	: **Yulianto, M.Pd.I** **NIP. 198707122019031005**	

Mengetahui,
Ketua Program Studi Perpustakaan dan Sains Informasi
Fakultas Sains dan Teknologi
Universitas Islam Negeri Maulana Malik Ibrahim Malang

Nita Siti Mudawamah, M.IP
NIP. 19[illegible]02232018012001

HALAMAN PERNYATAAN KEASLIAN TULISAN

Saya yang bertanda tangan dibawah ini:

Nama : Muhammad Dimas Alfathir

NIM : 220607110026

Prodi : Perpustakaan dan Sains Informasi

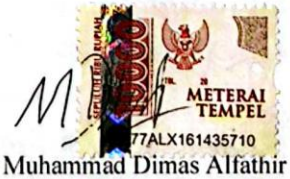
Fakultas : Sains dan Teknologi

Judul Skripsi : **Analisis Sentimen Komentar Penonton Youtube Terhadap Topik Korupsi Pada Kanal Edukasi Malaka Project Menggunakan Algoritma Random Forest**

Menyatakan dengan sebenarnya bahwa skripsi yang saya tulis ini benar-benar merupakan hasil karya saya sendiri, bukan merupakan pengambilan data tulisan atau pikiran orang lain yang saya akui sebagai hasil tulisan saya sendiri, kecuali dengan mencantumkan sumber terkait pada daftar pustaka.

Apabila di kemudian hari terbukti atau dapat dibuktikan skripsi ini hasil plagiasi, maka saya bersedia menerima sanksi atas perbuatan tersebut.

Malang, 22 April 2026



Muhammad Dimas Alfathir

NIM. 220607110026

KATA PENGANTAR

Assalamualaikum Warrahmatullahi Wabarakatuh.

Segala puji bagi Allah SWT tuhan semesta alam yang telah melimpahkan rahmat, hidayah, serta inayah-Nya, sehingga penulis dapat merampungkan penyusunan skripsi ini dengan judul "Analisis Sentimen Terhadap Komentar Penonton YouTube Terhadap Topik Korupsi Pada Kanal Edukasi Malaka Project Menggunakan Algoritma Random Forest". Tugas akhir ini disusun sebagai salah satu syarat akademik untuk meraih gelar Sarjana Sains Informasi (S.S.I) pada Program Studi Perpustakaan dan Sains Informasi, Fakultas Sains dan Teknologi, Universitas Islam Negeri Maulana Malik Ibrahim Malang. Shalawat serta salam semoga selalu terlimpahkan kepada baginda Nabi Muhammad SAW, sosok teladan yang telah membimbing umat manusia menuju jalan kebenaran dan peradaban yang mulia di bawah naungan Islam.

Peneliti menyadari bahwa selesainya skripsi ini tidak terlepas dari bantuan banyak pihak. Oleh karena itu, peneliti ingin mengucapkan terima kasih yang tulus kepada:

1. Kedua Orang Tua saya, khususnya mamah saya yang sebagai mamah pejuang yang selalu memberikan dukungan, berupa materi dan juga semangat motivasi untuk meyakini saya berkuliah hingga sampai saat ini. Saya ingin mengucapkan terima kasih atas kasih sayang yang tidak ada habis-habisnya diberikan kepada saya selama diperantauan. Semoga rezeki mamah selalu dilancarkan, diberikan kesehatan, serta kemudahan di setiap urusannya. Terimakasih sudah selalu mendukung dan memberikan rasa kepercayaan kepada anakmu ini.
2. Keluarga Besar Alm. Kakek Sunaryo dan Keluarga besar Kakek Rudi yang telah memberikan semangat dan dukungan selama masa perkuliahan, khsususnya saat semester ini.

3. Ibu Prof. Dr. Hj. Ilfi Nur Diana, M.Si., CAHRM., CRMP., selaku Rektor Universitas Islam Negeri Maulana Malik Ibrahim Malang.
4. Ibu Nita Siti Mudawamah, M.IP., selaku Ketua Program Studi Perpustakaan dan Sains Informasi.
5. Bapak Fakhris Khusnu Reza Mahfud, M. Kom. dan Bapak Ahmad Yulianto, M.Pd.I., selaku Dosen Pembimbing yang telah meluangkan waktu, memberikan arahan, serta bimbingan yang sangat berharga hingga skripsi ini dapat diselesaikan dengan baik.
6. Ibu Annisa Fajriyah, M.A dan Bapak Firma Sahrul Bahtiar, M.Kom., selaku Dosen Program Studi Perpustakaan dan Sains Informasi sekaligus penguji yang telah memberikan masukan, kritik, serta saran yang membangun demi kesempurnaan skripsi ini.
7. Bapak dan Ibu Dosen Jurusan Perpustakaan dan Sains Informasi, yang selalu memberikan ilmu dan pengalamannya selama mengajar di perkuliahan.
8. Untuk Rofiatul Laili, peneliti ingin mengucapkan terima kasih atas kesabaran, pengertian, dan dukungan yang diberikan kepada peneliti sampai proses penulisan skripsi ini selesai. Peneliti senang dan bersyukur memiliki kekasih yang baik hati. Semoga kebahagiaan dan kemudahan selalu membersamaimu, seperti halnya kamu selalu mendampingi peneliti.
9. Sahabat-sahabat seperjuangan selama perkuliahan, khususnya Aji, Dito, Fafa, Mizan, Iqbal, Zaka, Sultan, Heru, dan masih banyak sahabat-sahabat terbaik saya selama di perantauan, yang tidak bisa saya sebutkan satu persatu tetapi tanpa mengurangi rasa hormat saya. Semoga teman-teman saya ini diberikan kesehatan dan kelancaran dalam segala urusannya, terima kasih atas pengalaman yang baiknya selama di kota perantauan Malang.
10. Rekan-rekan di HMPS Libration 2022, terimakasih atas dukungan serta pengalamannya selama masa perkuliahan. Sukses selalu dan diberikan kelancaran dalam masa perkuliahannya.

11. Terakhir untuk saya sendiri, terima kasih sudah bertahan sejauh ini, tidak mudah menyerah dan putus asa pada kondisi, dan berani terus berkembang dalam setiap prosesnya. Sebelumnya sempat tidak ingin kuliah sampai berakhir di titik ini, “GUE BANGGA SAMA DIRI GUE SAAT INI”. Lanjutin cita-cita lu buat terus belajar sampai jadi dosen di umur 40 tahun.

Peneliti menyadari bahwa skripsi ini masih jauh dari kata sempurna. Oleh karena itu, kritik dan saran yang membangun sangat diharapkan. Semoga karya sederhana ini dapat memberikan manfaat bagi pembaca dan bagi perkembangan ilmu pengetahuan di bidang Sains Informasi.

Wassalamu’alaikum Warrahmatullahi Wabarakatuh.

Malang, 07 April 2025

Muhammad Dimas Alfathir

ABSTRAK

Alfathir, Muhammad Dimas. 2026. **Analisis Sentimen Komentar Penonton YouTube Terhadap Topik Korupsi Pada Kanal Edukasi Malaka Project Menggunakan Algoritma Random Forest. Skripsi. Program Studi Perpustakaan dan Sains Informasi Fakultas Sains dan Teknologi Universitas Islam Negeri Maulana Malik Ibrahim Malang. Pembimbing: (I) Fakhris Khusnu Reza Mahfud, M. Kom. (II) Ahmad Yulianto, M.Pd.I.**

Kata kunci: Analisis Sentimen, Random Forest, TF-IDF, Korupsi, Malaka Project.

Media sosial YouTube menjadi ruang digital bagi masyarakat untuk menyampaikan opini terhadap isu publik, termasuk korupsi. Penelitian ini bertujuan menganalisis klasifikasi sentimen komentar penonton pada dua video kanal edukasi Malaka Project, yaitu "Korupsi Terstruktur dan Terjahat Pertamina" dan "Vonis Tom Lembong Adalah Masalah Publik", serta membandingkan distribusi sentimen pada kedua video tersebut. Penelitian ini menggunakan pendekatan kuantitatif berbasis text mining dengan algoritma Random Forest. Data diperoleh melalui teknik scraping sebanyak 4.544 komentar, kemudian diberi label manual oleh tiga ahli bahasa ke dalam tiga kelas, yaitu positif, netral, dan negatif. Tahapan penelitian meliputi preprocessing text, pembobotan kata menggunakan TF-IDF, pembagian data latih dan data uji dengan rasio 80:20 dan 70:30, serta evaluasi model menggunakan confusion matrix. Hasil penelitian menunjukkan bahwa skenario rasio 70:30 menghasilkan performa terbaik dengan akurasi 76%, precision weighted average 72%, recall 76%, dan f1-score 73%. Namun, nilai macro average masih rendah, yaitu precision 54%, recall 46%, dan f1-score 48%, karena distribusi label tidak seimbang dan didominasi sentimen positif. Pada video pertama, sentimen positif mencapai 75%, sedangkan pada video kedua mencapai 78,6%. Proporsi sentimen negatif pada kasus korupsi individu hanya 3%, lebih rendah daripada korupsi sistemik sebesar 11%. Temuan ini menunjukkan bahwa penonton cenderung lebih defensif terhadap kasus korupsi individu. Penelitian ini menyarankan penggunaan teknik penanganan data tidak seimbang, perluasan sumber data, serta perbandingan algoritma lain agar deteksi sentimen minoritas lebih akurat dengan tetap memperhatikan konteks bahasa Indonesia dan dinamika percakapan digital.

ABSTRACT

Alfathir, Muhammad Dimas. 2026. **Sentiment Analysis of YouTube Viewers' Comments on the Topic of Corruption in the Malaka Project Educational Kanal Using the Random Forest Algorithm. Thesis. Library and Information Science Study Program, Faculty of Science and Technology, State Islamic University of Maulana Malik Ibrahim Malang. Advisors: (I) Fakhris Khusnu Reza Mahfud, M. Kom. (II) Ahmad Yulianto, M.Pd.I.**

Keywords: Sentiment Analysis, Random Forest, TF-IDF, Corruption, Malaka Project.

YouTube social media has become a digital space to express opinions on public issues, including corruption. The research aims to analyze the sentiment classification of viewer comments on two educational videos published by the Malaka Project channel, namely "Korupsi Terstruktur dan Terjahat Pertamina" and "Vonis Tom Lembong Adalah Masalah Publik". It also compares the sentiment distribution across the two videos. It employed a quantitative approach based on text mining using the Random Forest algorithm. The researcher collected 4,544 comments through scraping techniques and involved three language experts to manually label the comments into three categories—positive, neutral, and negative. The research procedures consisted of text preprocessing, term weighting using TF-IDF, splitting the dataset into training and testing sets with ratios of 80:20 and 70:30, and evaluating the model using a confusion matrix. The research results indicate that the 70:30 ratio scenario generated the best performance, achieving 76% accuracy, 72% weighted average precision, 76% recall, and a 73% f1-score. However, the macro average values remained low, with 54% precision, 46% recall, and a 48% f1-score, due to the imbalanced label distribution dominated by positive sentiment. In the first video, positive sentiment reached 75% of the comments, while in the second video it reached 78.6%. The proportion of negative sentiment in the case of individual corruption was only 3%, lower than the 11% observed in systemic corruption cases. These findings suggest that viewers tend to show more defensive attitude toward cases of individual corruption. Therefore, the research recommends the application of imbalance-handling techniques, the expansion of data sources, and comparisons with other algorithms to improve the accuracy of minority sentiment detection while still considering the Indonesian language context and the digital conversation dynamics.

مستخلص البحث

الفاطر، محمد ديماس. 2026. تحليل المشاعر لتعليقات مشاهدي يوتيوب حول موضوع الفساد في قناة مشروع مالاكا التعليمي باستخدام خوارزمية الغابة العشوائية. البحث الجامعي. قسم المكتبات وعلوم المعلومات، كلية العلوم والتكنولوجيا بجامعة مولانا مالك إبراهيم الإسلامية الحكومية مالانج. المشرف الأول: فخرس خُسن رضا محفوظ، الماجستير؛ المشرف الثاني: أحمد يوليانتو، الماجستير.

الكلمات الرئيسية: تحليل مشاعر، غابة عشوائية، TF-IDF، فساد، مشروع مالاكا.

أصبح وسائل التواصل الاجتماعي يوتيوب مساحة رقمية للمجتمع للتعبير عن آرائهم بشأن القضايا العامة، بما في ذلك الفساد. يهدف هذا البحث إلى تحليل تصنيف المشاعر لتعليقات المشاهدين على مقطعين فيديو لقناة مشروع مالاكا التعليمي، وهما "الفساد المنظم والأجرم لشركة فرتامينا" و"حكم توم لمبونج هو قضية عامة"، وكذلك مقارنة توزيع المشاعر بين هذين المقطعين. استخدم هذا البحث منهجا كميا مبنيا على استخراج النصوص باستخدام خوارزمية الغابة العشوائية. تم الحصول على البيانات من خلال تقنية الاستخراج وعددها 4544 تعليقًا، ثم تم وضع العلامات يدويًا من قبل ثلاثة خبراء لغويين إلى ثلاث فئات، وهي الإيجابية، والمحايدة، والسلبية. تشمل مراحل الدراسة معالجة النصوص المسبقة، وتوزين الكلمات باستخدام تي اف–اي دي دف (TF-IDF)، وتقسيم البيانات إلى بيانات تدريبية وبيانات اختبار بنسبة 80:20 و70:30، وكذلك تقييم النموذج باستخدام مصفوفة الارتباك .أظهرت نتائج البحث أن سيناريو النسبة 70:30 يعطي أفضل أداء بدقة 76٪، ومتوسط الوزن للثبات 72٪، والاسترجاع 76٪، ودرجة ف1 73%. ومع ذلك، فإن قيمة المتوسط الكلي لا تزال منخفضة، حيث الثبات 54٪، والاسترجاع 46٪، ودرجة ف1 48%، بسبب عدم توازن توزيع التصنيفات وسيطرة المشاعر الإيجابية. في الفيديو الأول، وصلت نسبة المشاعر الإيجابية إلى 75٪، بينما في الفيديو الثاني بلغت 78.6٪. ونسبة المشاعر السلبية في حالات الفساد الفردي هي فقط 3٪، وهي أقل من الفساد النظامي الذي يبلغ 11٪. تشير هذه النتائج إلى أن المشاهدين يميلون إلى اتخاذ موقف أكثر دفاعية في مواجهة حالات الفساد الفردية. يوصي هذا البحث باستخدام تقنيات معالجة البيانات غير المتوازنة، وتوسيع مصادر البيانات، وكذلك مقارنة الخوارزميات الأخرى من أجل جعل اكتشاف مشاعر الأقلية أكثر دقة مع مراعاة سياق اللغة الإندونيسية وديناميكيات المحادثات الرقمية.

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

BAB I

PENDAHULUAN

1.1 Latar Belakang

Perkembangan teknologi informasi menciptakan sebuah ruang digital untuk berinteraksi, di mana penggunanya dapat memperoleh informasi serta menyebarkan informasi, termasuk menyampaikan opini. Salah satu contohnya, yaitu media sosial yang berperan sebagai ruang digital oleh masyarakat dengan tujuan masyarakat dapat mengekspresikan opini, berbagi informasi, dan melakukan dialog interaktif (Utomo & Prayogi, 2021). Hal ini merubah pola perilaku masyarakat terhadap informasi, di mana masyarakat lebih mengutamakan kecepatan dalam memperoleh informasi dibandingkan dengan kebenaran sumber dan isi dari informasi tersebut.

YouTube menjadi salah satu media yang mendukung transformasi digital terhadap informasi. *Platform* ini tidak hanya menjadi tempat distribusi konten visual, tetapi juga memberikan kesempatan kepada pengguna untuk menggunggah, menonton, berinteraksi, dan membagikan video secara gratis. Hal ini membentuk ekosistem interaksi dua arah yang memungkinkan terjadinya komunikasi antara pembuat konten (*content creator*) dan penonton (*viewer*) melalui fitur interaktif seperti tombol *like,* kolom *comment*, dan *live chat* saat siaran langsung (Daraini et al., 2024). Saat ini, YouTube tidak hanya berperan sebagai media hiburan, tetapi berkembang menjadi sarana alternatif untuk media pembelajaran. Fenomena ini tergambar dari banyaknya kanal YouTube yang menyajikan konten-konten edukasi, baik dari sisi ilmu pengetahuan, peningkatan literasi informasi, atau pemahaman terkait isu-isu sosial.

Perubahan ini tidak hanya berdampak pada media yang digunakan, tetapi juga terlihat pada pola konsumsi informasi di tengah masyarakat. Dahulu, masyarakat cenderung mengandalkan media konvensional seperti percetakan sebagai sumber informasi. Namun, seiring dengan perkembangan teknologi informasi, mayoritas masyarakat beralih menggunakan *platform* media sosial

seperti YouTube, Instagram, dan berbagai *platform* lainnya sebagai sumber informasi utama. Di sisi lain, kehadiran ruang digital seperti YouTube memberikan kesempatan bagi masyarakat digital (*digital natives*) untuk berkontribusi dalam menyampaikan pandangan dan opini terkait isu yang berada di tengah masyarakat. Melalui fitur komentar masyarakat bebas berekspresi terhadap sebuah konten, seperti memberikan pujian, kritik tajam, atau ungkapan kekecewaan (Alviyanti & Said, 2025). Namun, kondisi tersebut memunculkan berbagai tantangan baru yang berkaitan dengan pelanggaran etika informasi di ruang digital, seperti penyebaran hoaks, tindakan *cyberbullying*, dan pelanggaran serupa lainnya.

Salah satu kanal YouTube yang secara konsisten menyajikan konten edukasi dan mendapatkan perhatian dari masyarakat yaitu Malaka Project. Malaka Project meruapakan proyek kolaboratif yang digagas oleh Ferry Irwandi bersama delapan *content creator* lainnya. Kehadiran Malaka Project bertujuan untuk memberikan kemudahan akses terhadap pendidikan melalui penyajian konten yang informatif dan kritis, sehingga dapat membentuk masyarakat baru yang memiliki tingkat literasi lebih baik dan terbiasa berpikir kritis dalam menghadapi berbagai permasalahan sosial sehari-hari (Nisrinaf, 2023).

Dalam penelitian ini, peneliti tidak membahas seluruh konten di kanal Malaka Project. Tetapi, memilih dua konten yang membahas topik “korupsi”. Pemilihan kedua konten ini dilatarbelakangi oleh tingginya jumlah kasus korupsi yang ditangani oleh Komisi Pemberantasan Korupsi (KPK) sepanjang 2020-2024, sebanyak 2.730 kasus yang tersebar pada lima sektor utama (KPK, 2024). Selain itu, topik korupsi cenderung menimbulkan beragam respon emosional dari penonton, seperti kemarahan, dukungan, atau ketidakpercayaan.

Konten pertama berjudul “Korupsi Terstruktur dan Terjahat Pertamina” yang membahas isu korupsi di sebuah perusahaan BUMN (Badan Umum Milik Negara). Alasan pemilihan video ini berkaitan dengan kasus korupsi terbaru dengan jumlah yang sangat besar, menurut Kompas kasus ini melibatkan tujuh

tersangka dalam dugaan kasus korupsi tata kelola minyak mentah dan produk kilang pada PT Pertamina yang mengakibatkan kerugian pada negara sebesar Rp. 193 Trilun. Hal tersebut menarik perhatian masyarakat secara luas, terutama berkaitan dengan menurunnya tingkat kepercayaan publik terhadap pemerintah maupun Badan Usaha Milik Negara (BUMN) (Nababan, 2025). Konten kedua berjudul "Vonis Tom Lembong Adalah Masalah Publik" yang membahas kasus impor gula yang dilakukan oleh mantan menteri perdagangan Tom Lembong dan dijatuhi vonis atas tindak pidana korupsi terkait kebijakan impor. Hal ini menarik perhatian masyarakat, karena banyaknya pejabat publik yang terbukti melakukan korupsi. Faktanya, pada kasus ini Tom Lembong tidak memperoleh keuntungan pribadi dari dugaan tindak pidana korupsi tersebut (Pitaloka, 2025).

Keberagaman topik yang terdapat pada kanal Malaka Project, terutama topik korupsi yang memicu berbagai respon dari penonton. Dalam konteks interaksi di media sosial, seperti penggunaan kolom komentar di *platform* YouTube, sentimen yang muncul dari penonton sangat beragam, mulai dari menyampaikan dukungan atau kritik tajam, hingga melayangkan ujaran kebencian dan penyebaran hoaks. Oleh karena itu, penting bagi masyarakat terutama pengguna media sosial untuk menyikapi informasi dan opini secara bijak. Dalam pandangan islam sikap selektif terhadap informasi telah ditegaskan dalam Qur'an Surat Al-Hujurat ayat 6 yang berbunyi:

يٰٓاَيُّهَا الَّذِيْنَ اٰمَنُوْٓا اِنْ جَاۤءَكُمْ فَاسِقٌۢ بِنَبَاٍ فَتَبَيَّنُوْٓا اَنْ تُصِيْبُوْا قَوْمًاۢ بِجَهَالَةٍ فَتُصْبِحُوْا عَلٰى مَا فَعَلْتُمْ نٰدِمِيْنَ ﴿٦﴾

Artinya: "*Wahai orang-orang yang beriman, jika seorang fasik datang kepadamu membawa berita penting, maka telitilah kebenarannya agar kamu tidak mencelakakan suatu kaum karena ketidaktahuan(-mu) yang berakibat kamu menyesali perbuatanmu itu."* (Q.S. Al-Hujurat ayat 6).

Ayat ini menekankan prinsip *tabayyun*, yakni melakukan klarifikasi terhadap setiap informasi sebelum mempercayai atau menyebarkannya. Pesan yang terkandung pada potongan ayat surah Al-Hujurat. Memiliki makna, jangan mudah perccaya pada orang yang diragukan kredibilitas nya dalam

mengonsumsi informasi. Dalam konteks beropini di media sosial, prinsip ini menjadi sangat relevan untuk membangun budaya literasi informasi yang etis dan bertanggung jawab. Oleh karena itu, diperlukan pendekatan yang mengintegrasikan prinsip etika informasi sekaligus mampu menganalisis kecenderungan opini publik secara sistematis (Qur'an Kemenag, 2022).

Penerapan analisis sentimen dipilih untuk memahami bagaimana opini publik terhadap kedua video dari topik korupsi. Pendekatan ini dinilai lebih objektif dan sistematis dalam menggambarkan keragaman respons yang muncul dari penonton. Analisis sentimen termasuk dalam bidang *Text Mining dan Natural Language Processing* (NLP). Proses ini dilakukan untuk menganalisis opini yang di ekspresikan ke dalam sebuah teks terhadap sebuah entitas, yang nantinya akan diklasifikasikan ke dalam beberapa label sentimen (Mahfud et al., 2020). Dalam pengoptimalisasi proses klasifikasi sentimen, diperlukan algoritma *machine learning* yang memiliki kemampuan dalam mengolah data teks dalam jumlah besar secara efektif, sehingga hasil prediksi yang didapatkan memiliki tingkat akurasi yang tinggi.

Dalam penelitian ini, algoritma *Random Forest* dipilih sebagai metode untuk mengoptimalkan hasil klasifikasi sentimen. Algoritma *Random Forest* merupakan salah satu teknik *ensemble learning*, dikenal karena memiliki tingkat akurasi prediksi yang tinggi serta kemampuannya dalam mengatasi data berukuran besar dan kompleks. *Random Forest* merupakan gabungan dari beberapa pohon keputusan (*Decision Tree*), di mana setiap pohon dibentuk dari subset acak data pelatihan, kemudian menggabungkan outputnya untuk meningkatkan kemampuan generalisasi model dengan proses *majority voting* (pemungutan suara terbanyak) (Ahmed et al., 2022).

Dalam penelitian terdahulu yang dilakukan oleh (Mario et al., 2025), yang berjudul "*Public Sentiment Analysis On Dirty Vote Movie On YouTube Using Random Forest and Naïve Bayes*". Penelitian ini menerapkan dua algoritma Random Forest dan Naïve Bayes untuk menganalisis sentimen terhadap film *Dirty Vote* di YouTube. Melalui tahapan *data crawling* dengan Apify, *text preprocessing*, *labelling data* dengan *library* VADER Sentiment dan

TextBlob, dan penerapan kedua algoritma Random Forest dan Naïve Bayes. Hasil menunjukan bahwa *Random Forest* dikombinasikan dengan label VADER Sentiment menghasilkan akurasi tertinggi sebesar 98,76% dan performa lebih baik dibandingkan Naïve Bayes. Penelitian tersebut menunjukkan bahwa Random Forest memiliki tingkat akurasi yang tinggi dan stabil dalam mengolah data berukuran besar dengan keragaman sentimen.

Pada penelitian ini berfokus terhadap penerapan analisis sentimen pada komentar YouTube di kanal edukasi Malaka Project, yang didukung oleh optimalisasi klasifikasi menggunakan algoritma *Random Forest*. Tahapannya dimulai mengumpulkan data komentar menggunakan teknik *Scraping*, kemudian dataset komentar dilabeli secara manual ke dalam tiga kategori sentimen, yaitu positif, negatif, dan netral dengan bantuan tiga orang ahli bahasa yang berprofesi sebagai guru di Yayasan Pondok Pesantren Nurul Islam Mojokerto. Selanjutnya *text preprocessing*, menggunakan *library* Sastrawi untuk mengolah data teks berbahasa Indonesia, selanjutnya tahap pembobotan kata menggunakan metode TF-IDF (*Term Frequency – Inverse Document Frequency*). Selanjutnya, dilakukan pembagian dataset untuk data latih dan data uji yang digunakan dalam pelatihan algoritma. Penerapan algoritma *Random Forest* untuk menghasilkan prediksi dari klasifikasi sentimen pada dataset yang sudah bersih dan konsisten (Ahmed et al., 2022). Mengevaluasi kinerja algoritma dengan tabel *confusion matrix* ditinjau berdasarkan hasil *accuracy*, *precision*, *recall*, dan *f1-score* (Ramadhani & Suryono, 2024). Penelitian ini diharapkan dapat memberikan *insight* bagi *content creator* mengenai persepsi publik terhadap isu-isu sensitif, khususnya dalam ketiga kategori, yaitu sosial-politik, kontroversial, dan edukasi yang disajikan melalui media sosial, terutama *platform* YouTube.

1.2 Identifikasi Masalah

1. Bagaimana hasil klasifikasi sentimen pada komentar penonton terhadap topik korupsi di kanal YouTube edukasi Malaka Project

menggunakan algoritma *Random Forest* yang ditinjau berdasarkan nilai *accuracy, precision, recall,* dan *f1-score*?

2. Bagaimana perbandingan distribusi sentimen (positif, negatif, dan netral) pada komentar penonton antara video 'Korupsi Terstruktur dan Terjahat Pertamina' dan video 'Vonis Tom Lembong Adalah Masalah Publik'?

1.3 Tujuan Penelitian

1. Untuk menganalisis hasil klasifikasi sentimen pada komentar penonton terhadap topik korupsi di kanal YouTube edukasi Malaka Project menggunakan algoritma *Random Forest* yang ditinjau berdasarkan nilai *accuracy*, *precision*, *recall*, dan *f1-score*.
2. Untuk menganalisis perbandingan distribusi sentimen (positif, negatif, dan netral) pada komentar penonton antara video “Korupsi Terstruktur dan Terjahat Pertamina” dan video “Vonis Tom Lembong Adalah Masalah Publik”.

1.4 Manfaat Penelitian

Hasil penelitian ini diharapkan dapat memperkaya kajian dalam bidang Sains Informasi, khususnya dalam penerapan data mining dan analisis sentimen berbasis teks pada media sosial seperti YouTube. Selain itu, penelitian ini dapat memberikan gambaran mengenai opini yang tersebar di kolom komentar mengenai isu-isu sensitif, seperti korupsi. Di mana insight dari pola sentimen bisa dimanfaatkan oleh *content creator* sebagai strategi pengembangan konten dari sisi komunikasi digital melalui media sosial.

1.5 Batasan Masalah

Adapun batasan dalam penelitian untuk menghindari perluasan pembahasan sehingga tidak terfokus, perlu adanya batasan dalam penelitian. Batasan masalah pada penelitian ini, terdiri dari:

1. Objek penelitian terbatas pada topik korupsi dari video “Korupsi Terstruktur dan Terjahat Pertamina” dan “Vonis Tom Lembong Adalah Masalah Publik”.
2. Waktu pengambilan data komentar dilakukan pada 20 Oktober 2025.

1.6 Sistematika Penulisan

Sistematika penulisan dibuat agar mempermudah pembaca dalam memahami alur penulisan pada penelitian ini, sistematika terdiri dari lima bab yaitu:

BAB I Pendahuluan

Pada bab ini membahas latar belakang penelitian yang menjelaskan transformasi informasi di era digital, khususnya peran media sosial (YouTube), serta penggunaan analisis sentimen dalam memahami opini publik dan integrasi nilai-nilai keislaman. Bab ini juga memuat identifikasi masalah, tujuan, dan batasan mengenai penelitian tentang analisis sentimen terhadap komentar penonton YouTube pada kanal edukasi Malaka Project menggunakan algoritma *Random Forest* yang ditinjau dari nilai akurasi, presisi, recall, dan F1-score.

BAB II Tinjauan Pustaka

Pada bab ini berisi lima penelitian terdahulu yang relevan dengan topik analisis sentimen pada media sosial, serta menyajikan kerangka teori yang menjadi landasan penelitian. Pada bagian landasan teori membahas secara mendalam konsep-konsep yang mendukung penelitian ini, seperti *text mining*, analisis sentimen, Malaka Project, tahapan *preprocessing text*, pembobotan kata (TF-IDF), dan algoritma *Random Forest* sebagai model klasifikasi sentimen yang dievaluasi melalui metrik *accuracy*, *precision*, *recall*, dan *f1-score*. Terdapat sub-bab integrasi keislaman berdasarkan ayat-ayat dan tafsir dari Al-Qur’an sebagai dasar etika berkomentar di media sosial.

BAB III Metodologi Penelitian

Pada bab ini menjelaskan metode penelitian yang digunakan, meliputi jenis penelitian yaitu penelitian kuantitatif, selanjutnya alur penelitian merupakan tahapan-tahapan dalam penelitian dari awal data di ambil sampai penarikan kesimpulan, kemudian desain sistem untuk mencapai tujuan penelitian melalui penggunaan aplikasi yang mendukung proses analisis sentimen, selanjutnya sumber data yang diperoleh dari situs web, kemudian data dilakukan tahapan pra-pemrosesan teks, dan evaluasi hasil dari penerapan algoritma *Random Forest* menggunakan *confusion matrix*.

BAB IV Hasil dan Pembahasan

Pada bab ini menyajikan hasil dari proses analisis sentimen yang telah dilakukan, dimulai dari tahapan *preprocessing*, visualisasi data, penerapan algoritma Random Forest, evaluasi *confusion matrix*, dan visualisasi *wordcloud*. Bab ini merepresentasikan kecenderungan sentimen masyarakat terhadap konten edukasi pada kanal Malaka Project, serta analisis terhadap hasil klasifikasi.

BAB V Kesimpulan dan Saran

Pada bab ini merangkum hasil penelitian, menarik kesimpulan mengenai efektivitas algoritma Random Forest dalam mengklasifikan sentimen, serta memberikan saran bagi penelitian lanjutan dan pengembangan strategi komunikasi oleh content creator dalama menyajikan konten edukatif di media sosial.

BAB II
TINJAUAN PUSTAKA

2.1 Tinjauan Pustaka

Berbagai penelitian terdahulu telah mengeksplorasi penerapan analisis sentimen pada komentar media sosial dengan menggunakan beberapa algoritma klasifikasi. Namun, belum ada kajian secara eksplisit yang menerapkan dalam konteks edukatif seperti di *kanal YouTube* Malaka Project. Adapun beberapa penelitian terdahulu yang melandasi penelitian ini, seperti penelitian berjudul "*Analisis Sentiment Masyarakat terhadap Kasus Covid-19 pada Media Sosial YouTube dengan Metode Naïve Bayes*" menjelaskan perkembangan kasus Covid-19 di Indonesia yang terus meningkat, berdasarkan laman Covid-19 nasional rata-rata kenaikan kasus aktif dari bulan april sampai desember tahun 2020 sebesar 8.894 kasus. Penelitian ini bertujuan membangun model untuk mengklasifikasikan komentar masyarakat pada pemberitaan tentang Covid-19 di *kanal* KompasTV. Pengklasifikasian sentimen terbagi ke dalam tiga kategori, yaitu positif, negatif, dan netral yang menggambarkan persepsi publik terhadap informasi pandemi Covid-19. Data komentar diperoleh melalui teknik *scraping*-komentar menggunakan ekstensi Data Miner di Chrome, kemudian melalui tahapan *pre-processing*, dengan tokenisasi, pembersihan data, stemming, dan transformasi huruf. Selanjutnya pemberian label sentimen berdasarkan *score* kamus sentiment (positif, negatif, netral). Pembobotan kata dilakukan dengan TF-IDF sebelum pelatihan model klasifikasi, selanjutnya menggunakan algoritma *Naïve Bayes* dan divalidasi menggunakan *K-Fold Cross Validation*. Hasil penelitian ini, menggambarkan persepsi publik mayoritas sentimen negatif, dengan total komentar 800 negatif dibandingkan 361 komentar positif dan 490 netral. Dengan tingkat akurasi model *Naïve Bayes* sebesar 74%. Kesimpulannya respon publik terhadap pemberitaan Covid-19 ini cenderung negatif terhadap pemberitaan Covid-19, serta kinerja algoritma *Naïve Bayes* cukup memuaskan terhadap data tersebut (Ahmadi et al., 2021).

Penelitian selanjutnya, yang berjudul "*Analisis Sentimen Komentar Pada Saluran YouTube Beauty Vlogger Berbahasa Indonesia Menggunakan Metode Support Vector Machine*" menjelaskan *YouTube* sebagai *platform* terbesar sebagai media distribusi video dengan cakupan *global* dan memberikan kebebasan kepada pengguna maupun perusahaan untuk menyebarkan konten edukatif, hiburan, hingga informatif terhadap suatu produk. Dalam dunia kecantikan, *platform* ini menjadi sarana utama bagi *Beauty Vlogger* memberikan rekomendasi, promosi, dan mengulas terhadap suatu produk kecantikan. Hal ini menunjukkan pentingnya ulasan pada komentar penonton dalam membangun persepsi dan keputusan konsumen terhadap suatu produk sebelum dibeli. Penelitian ini bertujuan menganalisis sentimen komentar berbahasa Indonesia pada ulasan bedak dan *skincare* di saluran *Beauty Vlogger* menggunakan algoritma *Support Vector Machine* dengan mengubah kata dengan TF-IDF (*Term Frequency-Inverse Document Frequency*), serta evaluasi kinerjanya dengan metrik *accuracy*, *precision,* dan *recall*. Data komentar sebanyak 1.000 komentar di *scraping*, kemudian pembagian dataset 800 untuk data latih dan 200 untuk data uji. Selanjutnya diproses dengan tahapan data *pre-processing* seperti tokenisasi, penghapusan *stopword*, dan *stemming* sebelum dilanjutkan ke tahap merubah kata menjadi vektor numerik dengan TF-IDF. Algoritma SVM (*Support Vector Machine*) dilatih dengan dataset yang sudah di bagi, lalu dievaluasi dengan ketiga metrik yang menghasilkan tingkat *accuracy* sebesar (97%), *precision* (98%), dan *recall* (96%), menunjukkan algoritma ini sangat tepat digunakan dalam mendeteksi sentimen positif dan negatif pada komentar berbahasa Indonesia (Adek et al., 2025).

Penelitian lainnya tentang analisis sentimen menggunakan algoritma *Random Forest* yang berjudul "*Penerapan Algortitma Random Forest Untuk Analisis Sentimen Komentar di YouTube Tentang Islamofobia*" menjelaskan ekspresi kecemasan individu maupun kelompok sosial terhadap agama Islam dan komunitas muslim yang muncul sebagai respon atas peristiwa pengeboman dan teror yang mengatasnamakan Islam. Islamofobia di internet merupakan

kekerasan verbal, sementara sentimen yang diperoleh dari media *platform* berbagi video seperti *YouTube* sangat beragam. Oleh karena itu, penelitian ini menerapkan pendekatan *text mining* berbasis pembelajaran mesin agar lebih efektif dan efisien dalam menganalisis teks. Tujuannya untuk memberikan edukasi dan pemahaman terhadap pelaku perundungan atau pelaku ujaran kebencian mengenai agama Islam di Indonesia. Tahapannya melalui pengumpulan data dengan dua tahapan yaitu *crawling* dan *labelling*, pengambilan data komentar *YouTube* menggunakan *ScrapeStorm* sebanyak 1.000 komentar. Selanjutnya tahap *labelling* dilakukan secara manual, dengan mempertimbangkan komentar-komentar yang mengandung kalimat Islamofobia, termasuk *hate speech* ataupun ujaran kebencian terhadap agama Islam dikategorikan sebagai label negarif. Sedangkan kalimat yang mengandung dukungan, ucapan baik, dan opini pembelaan Islam dikategorikan label positif. Selanjutnya tahapan teks *pre-processing* meliputi *Cleaning* untuk memberishkan *noise* (emoji, angka, tanda baca, dan lainnya), *Case Folding* untuk penyeragaman huruf kecil (*lowercase*), *Tokenizing* untuk memecah kata pada kalimat, *Stopword Removal* untuk menghapus kata yang tidak memiliki arti dalam kalimat, dan *Stemming* untuk menemukan kata dasar menggunakan pustaka *Sastrawi* untuk teks bahasa Indonesia. Selanjutnya langkah pembobotan kata menggunakan TF-IDF (*Term Frequency-Inverse Document Frequency*) untuk menghitung kemunculan setiap kata didalam dokumen.

Setelah pembobotan, dilakukan pembangunan model klasifikasi dengan algoritma *Random Forest* untuk mengenal pola sebuah data untuk diklasifikasikan kedalam dua kelas, yaitu kelas positif dan negatif. Beberapa hiperparameter dalam model *Random Forest* juga di eksplorasi antara lain *n_estimators*, *max_depth*, dan *min_samples_split*. Pada penelitian ini dilakukan beragam *splitting data* untuk data latih data uji, digunakan rasion 70:30, 80:20, dan 90:10 bertujuan untuk menilai performa model pada ukuran data uji yang berbeda. Kinerja model dievaluasi dengan metrik *accuracy*, *precision*, *recall*, dan *F1-score* yang dihitung berdasarkan rumus *confusion matrix*. Hasilnya diperoleh akurasi tertinggi didapatkan sebesar 79% pada

pembagian data latih 90:10. Dengan kombinasi hiperparameter *n_estimators* = 10, *max_depth* = 25, dan, *min_samples_split* = 10. Berdasarkan rangkaian pengujian model yang digunakan untuk mengklasifikasikan teks menunjukkan lebih banyak komentar positif dibandingkan komentar negatif. Dengan akurasi sebesar 79%, presisi 79%, recall 95%, dan F1-Score 86,26 % yang merepresentasikan algoritma *Random Forest* cukup baik dalam mengklasifikasikan sentimen pada komentar *YouTube* dengan topik Islamofobia (Afdhal et al., 2022).

Penelitian selanjutnya berjudul "*Analisis Sentimen Publik pada Film Dirty Vote di YouTube Menggunakan Random Forest dan Naïve Bayes*" menjelaskan kehadiran *YouTube* sebagai media distribusi video digital yang memungkinkan pembuat film berinteraksi langsung dengan penonton dan membentuk persepsi dalam merespon sebuah karya audiovisual. Salah satunya, film *Dirty Vote* yang memicu perdebatan terkait isu masalah integritas dan manipulasi suara di tengah pemilu presiden dan wakil presiden tahun 2024. Mengidentifikasi sentimen positif atau negatif dari penonton, peneliti menggunakan pendekatan *text mining* dengan menggunakan algoritma *Random Forest* dan *Naïve Bayes*. Penelitian ini memberikan pemahaman mengenai respon penonton dan menguji kedua algoritma tersebut dalam mengidentifikasi sentimen pada film *Dirty Vote*. Tahapannya mengumpulkan data komentar *YouTube* dengan *website Apify* sebanyak 4.551 komentar, selanjutnya data yang sudah tersimpan dalam format CSV dilakukan pra-pemrosesan yang meliputi *Data Cleaning* untuk membersihkan *noise*, seperti (tanda baca, URL, data duplikat, dan data kosong), *Case Folding* dilakukan untuk merubah huruf dalam teks menjadi huruf kecil (*lowercase*), *Tokenizing* untuk membagi teks ke dalam token kecil, seperti frasa, kata atau simbol, *Filtering* untuk menyaring kata-kata yang tidak memiliki makna, dan *Stemming* untuk merubah kata menjadi bentuk asli dengan menghilangkan imbuhan awal, tengah, dan akhir kata.

Selanjutnya tahapan *labelling* dilakukan dengan menggunakan pustaka *vadeSentiment* dan *TextBlob*, dengan dua bahasa: Indonesia dan Inggris untuk menghasilkan label positif dan negatif. Selanjutnya mengekstraksi fitur dengan

TF-IDF untuk menghitung frekuensi kata yang muncul dalam satu dokumen, dan penerapan algoritma klasifikasi pertama *Random Forest* merupakan algoritma penggabungan pohon keputusan untuk menghasilkan model prediksi berdasarkan kombinasi pohon-pohon keputusan yang ditentukan dengan *majority voting* (mayoritas suara terbanyak), algoritma kedua *Naïve Bayes* merupakan algoritma perhitungan probabilitas, yang didasari dari distrbusi probabilitas pada fitur-fiturnya. Tahap selanjutnya evaluasi model dengan *classification report* (akurasi, presisi, recall, dan F1-score) untuk masing-masing bahasa (Indonesia dan Inggris). Pada algoritma *Random Forest* memperoleh hasil lebih baik dibanding *Naïve Bayes*, untuk dataset bahasa Indonesia dengan *TextBlob* memperoleh akurasi 97,98% recall positif meningkat 20% dari *Naïve Bayes* menjadi 28%, dalam *TextBlob* bahasa Inggris akurasi 81,01% recall positif 66 dan F1-Score positif lebih baik 0.73 dibanding *VADER* 0.64. Dari berbagai kombinasi *Random Forest* lebih unggul, namun *Naïve Bayes* menjadi alternatif yang efisien untuk dataset bahasa Indonesia. Namun, keunggulan *Random Forest* juga membutuhkan sumber daya komputasi yang lebih berat dan waktu pemrosesan yang lama dibandingkan dengan *Naïve Bayes* yang lebih cepat dalam pemrosesan data tetapi memiliki kelemahan dalam menangani dataset dengan fitur yang saling bergantung. Oleh karena itu, pemelihan algoritma harus disesuaikan dengan kebutuhan analisisnya (Mario et al., 2025).

Penelitian terakhir yang berjudul "*Public Opinions on ChatGPT: An Analysis of Reddit Discussions by Using Sentiment Analysis, Topic Modeling, and SWOT Analysis*" menjelaskan kemunculan *ChatGPT* salah satu produk *Artificial Intelligence* (AI) pada tanggal 30 November 2022 oleh perusahaan teknologi *OpenAI*. Fenomena ini sangat kontroversial di seluruh dunia, dalam waktu lima hari *ChatGPT* berhasil mendapat satu juta pengguna. Penggunanya sangat kagum dengan penemuan teknologi ini karena kemampuannya dalam menyelesaikan masalah sangat tepat dan cepat. Karena pertumbuhan pengguna yang sangat cepat, peneliti ingin mendalami opini para penggunanya terhadap produk *ChatGPT* di media sosial *Reddit*, rentang waktu pengambilan data yang

dilakukan selama satu tahun dari tahun 2022 sampai 2023. Peneliti bertujuan memahami sudut pandang masyarakat terhadap *ChatGPT*, serta mengidentifikasi potensi dan kekhawatiran yang muncul dari diskusi di media sosial *Reddit*. Peneliti menentukan objek dari situs web, khususnya dari komunitas *ChatGPT* di *Reddit*. Pengambilan data pada bulan Desember 2022 sampai Desember 2023, kemudian proses ekstraksi menggunakan pemograman *Python* dan *Reddit API* mendapatkan total 202.905 komentar. Peneliti melakukan tahapan *filtering* pada komentar yang tidak berbahasa inggris, setelah selesai total komentar yang siap untuk dianalisis lebih lanjut sebanyak 164.900 komentar. Selanjutnya, pada tahapan data *pre-processing* peneliti melakukan *lower casing* untuk merubah semua teks komentar menjadi huruf kecil agar lebih konsisten. Dilanjutkan tahap *stopwords and noise removal* untuk menghapus kata-kata umum yang sering muncul tapi tidak memiliki makna penting dalam teks dan membersihkan tanda baca seperti tagar, *URL*, dan emoji. Kemudian, dilakukan tahap *duplication and missing values removal* untuk menghapus komentar yang sama (duplikat) dan menghapus baris teks yang kosong agar dataset tetap konsisten.

Tahapan terakhir *lemmatization* untuk merubah kata menjadi bentuk dasar tujuannya mengurangi variasi kata. Setelah, tahap *pre-processing* dilakukan tahap *Topic Modelling* untuk menemukan topik apa saja yang sedang dibicarakan oleh anggota komunitas *ChatGPT* di *Reddit*. Metode yang digunakan adalah LDA (*Latent Dirchlet Allocation*) di mana prosesnya menghitung *coherence score* untuk menentukan jumlah topik dari data komentar. Setelah di tentukan, model LDA dilatih dengan data tersebut dan hasilnya mengekstrak kata-kata kunci yang paling relevan untuk setiap topik, selanjutnya peneliti mengoreksi manual komentar acak dari setiap topik untuk penamaan pada topik tersebut. Pada tahap selanjutnya peneliti melakukan analisis sentimen dengan membagi tiga klasifikasi sentimen (positif, negatif, dan netral) peneliti menerapkan enam model terdiri dari LogisticRegression, Naïve Bayes, Support Vector Machine (SVM), Random Forest, VADER, dan TextBlob. Didapatkan hasil dari model terbaik yaitu VADER (*Valence Aware*

Dictionary and Sentiment Reasoner) dengan akurasi 93,2%. Tahapan terakhir, melakukan analisis SWOT (*Strengths, Weakness, Opportunities, Threats*) terhadap *ChatGPT* berdasarkan opini pengguna di *Reddit*. Hasil analisis ini untuk pertimbangan bagi pengembang produk *ChatGPT* maupun pembuat kebijakan terkait AI (Naing & Udomwong, 2024).

Dari lima penelitian terdahulu yang telah di jelaskan sebelumnya, beberapa penelitian mempunyai kesamaan pada topik dan algoritma yang digunakan. Namun, terdapat perbedaan terhadap objek penelitian, di mana penelitian saat ini berfokus kepada sebuah kanal edukasi dengan topik korupsi, dibandingkan dengan penelitian terdahulu yang berfokus kepada isu-isu umum.

2.2 Landasan Teori

Pada sub bab ini menyusun landasan teori sebagai bentuk kerangka konseptual yang menjadi dasar bagi pemecehan masalah yang telah diidentifikasi. Pembahasan dimulai dengan konsep dasar *Text Mining* dan *Analisis Sentimen* sebagai metodologi utama untuk mengekstraksi data komentar tidak terstruktur. Tahapan teknis dijelaskan lebih mendalam melalui *Preprocessing Text* dan Pembobotan Kata *TF-IDF* untuk merubah teks menjadi format numerik yang memiliki nilai agar mudah diolah menggunakan sistem. Selanjutnya, dijelaskan prinsip kerja algoritma *Random Forest* sebagai metode klasifikasi, beserta *Confusion Matrix* sebagai instrumen evaluasi model. Serta disajikan profil Malaka Project, dengan integrasi nilai keislaman melalui prinsip *Tabayyun* dan *Fikih Perkataan* sebagai landasan etika komunikasi di ruang digital.

2.2.1 Text Mining

Text mining secara luas dapat diartikan proses intensif pengetahuan yang memungkinkan pengguna berinteraksi dengan kumpulan dokumen dari waktu ke waktu menggunakan seperangkat alat analisis. Berbeda dengan data mining tradisional, metode ini lebih kompleks karena bekerja

dengan kumpulan data yang tidak terstruktur, sehingga memerlukan tahapan *preprocessing* yang mencakup luas karena teks merepresentasikan pemrosesan secara komputasional (Hassani et al., 2020). Proses ini merupakan gabungan dari berbagai disiplin ilmu seperti *Natural Language Processing* (NLP), *Machine Learning*, *Information Retrieval System*, dan *Knowledge Management* untuk mengekstraksi pola-pola dari sekumpulan dokumen.

Dalam merepresentasikan dokumen, metode *text mining* menggunakan model yang memilih sebagian fitur untuk mewakili isi dokumen secara keseluruhan. Fitur yang umum digunakan seperti karakter, kata, istilah (*term*), dan konsep. Adapun tantangan utama yang dihadapi dalam *text mining* adalah jumlah fitur yang banyak serta fitur yang jarang muncul (*feature sparsity*) (Hassani et al., 2020).

2.2.2 Analisis Sentimen

Analisis sentimen atau *opinion mining* merupakan bidang studi yang mempelajari opini, penilaian, sikap, dan emosi yang diungkapkan terhadap suatu entitas beserta atributnya. Analisis sentimen berfokus pada mendeteksi orientasi (positif, negatif, atau netral) dan intensitas perasaan yang mendasari pernyataan, termasuk konteks emosionalnya. Terdapat dua tingkat analisis sentimen, yaitu tingkat dokumen dan tingkat kalimat. Pada tingkat dokumen bertujuan untuk mengklasifikasikan apakah dokumen berisi opini positif atau negatif. Kelebihannya adalah permodelan yang sederhana seperti klasifikasi teks umum karena hanya ada dua kelas, kelemahannya terdapat kekurangan pada kebutuhan analisis yang lebih mendalam terkait “bagian mana yang disukai/tidak disukai”. Selanjutnya pada tingkat kalimat untuk menentukan sebuah kalimat merepresentasikan ungkapan positif, negatif, atau netral (opini netral biasanya berarti tidak ada pendapat). Pada tahap ini tantangan muncul karena kalimat objektif menimbulkan kesan makna tersirat, sementara

kalimat interogratif, kondisional, dan sarkasme dapat memuat sentimen yang memiliki makna berbeda (Ardiani et al., 2020).

Dengan mengetahui tingkatan pada analisis sentimen, fokus analisis berlanjut ke tahapan mengukur dan mengklasifikasikan dokumen teks secara efektif dari data teks tidak terstruktur. Terdapat dua pendekatan umum yang digunakan dalam analisis sentimen, yaitu pendekatan berbasis *machine learning* dan pendekatan berbasis *lexicon-based approach*. Pendekatan berbasis *machine learning* menerapkan sebuah model komputasi yang dilatih untuk mengatasi masalah klasifikasi teks agar dapat membedakan antara teks positif dan negatif. Tahapannya, sebuah model diberikan data dalam jumlah besar yang sudah diberi label sentimen secara manual (data latih). Kemudian, dari data ini model akan belajar mengenali pola yang menjadi ciri khas setiap sentimen. Beberapa fitur yang digunakan untuk mengenali pola tersebut, seperti *Term Frequency* untuk menghitung intensitas kata yang muncul dalam teks yang nantinya akan diolah oleh algoritma *machine learning*. Sementara itu, berbeda dengan pendekatan berbasis *lexicon-based Approach* menggunakan kamus yang berisi daftar kata-kata beserta skor polaritasnya (positif dan negatif). Prosesnya analisisnya sederhana, seperti sistem mengidentifikasi kata-kata sentimen dalam sebuah kalimat dengan memperhatikan kata-kata pembalik seperti “tidak” (sentiment shifters), selanjutnya memprediksi skor total dalam menentukan sentimen keseluruhan dari kalimat tersebut. Keunggulan pendekatan ini tidak membutuhkan data latih. Namun, tantangannya terdapat pada memahami kata-kata sentimen yang bergantung pada konteks, serta kekurangannya memahami opini implisit dan sarkasme (Anam et al., 2023).

2.2.3 Preprocessing Text

Preprocessing text merupakan tahapan awal dalam alur kerja *Natural Language Processing* (NLP) yang memiliki peran penting dalam menentukan kualitas hasil akhir dari analisis sentimen. Tahapan ini

bertujuan untuk membersihkan dan mempersiapkan data teks mentah yang tidak terstruktur menjadi format teks yang lebih konsisten sehingga mudah diolah oleh algoritma. Data teks mentah biasanya berisi *noise* (kesalahan ejaan, penggunaan simbol, atau kata-kata kurang pantas) fenomena ini merupakan tantangan dalam proses analisis dan berdampak pada akurasi model *machine learning* yang sedang dibangun. (Raya et al., 2025). Tahapan ini melibatkan serangkaian proses yang diterapkan secara berurutan untuk memastikan data bersih dari *noise* dan konsisten. Secara umum rangkaian tersebut terdiri dari *case folding*, *text cleaning*, *normalization*, *tokenization*, *stopword removal*, dan *stemming leamtization* (Aufar et al., 2023).

a. Case Folding

Case folding merupakan proses menyeragamkan seluruh teks menjadi huruf kecil (*lowercase*). Teknik ini sangat efektif untuk mengatasi masalah inkonsistensi teks yang disebabkan keseragaman penggunaan huruf kapital (Raya et al., 2025). Dengan melakukan tahapan ini, kata bervariasi seperti “BAGUS”, “Bagus”, dan “bagus” akan dianggap sebagai kata yang sama oleh sistem, sehingga mengurangi kata-kata yang berlebihan dengan makna yang sama dalam proses analisis (Aufar et al., 2023).

b. Text Cleaning

Tahapan selanjutnya setelah menyeragamkan teks, adalah pembersihan teks atau biasa disebut proses *text cleaning*. Proses ini berfokus pada menghilangkan berbagai elemen-elemen yang tidak relevan pada teks dalam analisis sentimen. Elemen-elemen yang terdiri dari tanda baca (koma, titik, tanda tanya), angka, simbol khusus (@,#,&), emoji, hashtag, mention, serta tautan URL (Anggreny, 2025). Tahapan ini bertujuan menghasilkan dataset yang jauh lebih bersih dan seragam, sehingga model *machine learning* dapat lebih fokus ke teks yang mengandung sentimen (Raya et al., 2025).

c. Normalization

Normalisasi adalah tahapan mengubah kata-kata yang tidak baku, seperti bahasa gaul (*slang word*), singkatan, atau kesalahan ketik, merubahnya ke dalam bentuk kata baku yang sesuai dengan kaidah dalam KBBI (Kamus Besar Bahasa Indonesia) untuk bahasa Indonesia, misalnya kata "yg" diubah mejadi kata "yang" dan kata "ga" menjadi kata "tidak". Tahapan ini bertujuan untuk memastikan teks dapat diproses secara konsisten dan akurat oleh model, karena variasi kata yang tidak standar dapat mengganggu proses analisis (Anggreny, 2025).

d. Tokenization

Tokenization adalah tahapan pemecahan teks menjadi token-token kata yang lebih kecil. Token ini berupa kata individual, frasa, dan tanda baca, tergantung pada tujuan analisis (Aufar et al., 2023). Sebagai contoh, pada kalimat "program ini sangat membantu" akan dipecah menjadi token-token seperti ("program", "ini", "sangat", "membantu"). Pada tahap ini merupakan awal untuk analisis pada level kata, yang mempengaruhi sistem komputasi memahami dan memproses teks dengan lebih efisien (Anggreny, 2025).

e. Stopword Removal

Tahapan ini dilakukan untuk mengidentifikasi dan menghilangkan kata-kata umum yang sering muncul tetapi tidak relevan terhadap sentimen teks. Misalnya, kata-kata seperti "di", "yang", "atau", "dari", dan lainnya (Raya et al., 2025). Dengan menghilangkan kata-kata ini analisis lebih terfokus pada kata-kata yang memiliki makna sentimen, sehingga kualitas dan relevansi data menjadi lebih akurat (Aufar et al., 2023).

f. Stemming

Tahapan dalam menyederhanakan teks memiliki dua teknik umum, yaitu *stemming* dan *lemmatization*. *Stemming* dilakukan untuk merubah kata-kata yang memiliki imbuhan menjadi bentuk dasar (*root form*) dengan cara menghapus kata awalan, sisipan, atau akhiran pada teks

(Aufar et al., 2023). Misalnya, proses ini dilakukan pada kalimat "jaringan" dan "jaringannya", hasilnya dalam bentuk dasar "jaring". Proses ini sangat berguna untuk mengurangi keseragaman makna pada fitur dan meningkatkan konsistensi data yang digunakan pada model (Raya et al., 2025).

2.2.4 Pembobotan Kata: TF-IDF

Dalam alur kerja setelah tahapan *preprocessing text*, tahap selanjutnya mengubah teks menjadi format numerik (pembobotan kata) agar mempermudah pemrosesan oleh algoritma *machine learning*. Salah satu metode yang umum digunakan untuk pembobotan kata adalah TF-IDF. *Term Frequency-Inverse Document Frequency* (TF-IDF) atau *term weighting* merupakan salah satu metode dalam pengolahan teks dan membentuk model NLP (*Natural Language Processing*). Metode ini digunakan untuk mengukur substansi kata penting (*term*) terhadap kumpulan dokumen yang lebih besar (Agustina et al., 2024). Tujuan utama dalam metode ini adalah memberikan bobot pada setiap kata, di mana bobot tersebut merepresentasikan signifikasi kata pada seluruh dokumen. Metode ini membantu dalam mengidentifikasi kata-kata yang relevan pada suatu dokumen. Perhitungan bobot ini berdasarkan dua factor, yaitu *Term Frequency* dan *Inverse Documen Frequency* (Septiani & Isabela, 2022).

a. Term Frequency (TF)

Term Frequency digunakan untuk menghitung kemunculan setiap *term* dalam setiap dokumen. Setiap kata yang sering muncul pada suatu dokumen, menunjukkan kata itu sangat merepresentasikan isi dari dokumen tersebut (Pranata et al., 2024). Dalam penerapannya, sebuah rumus dijelaskan mengenai penghitungan TF ini, rumusnya adalah:

$$\boldsymbol{TF(t,d) = f_{t,d}} \tag{2.1}$$

Keterangan:

TF (t,d): Nilai Term Frequency dari sebuah kata t di dalam sebuah dokumen d

t: Kata (*term*) spesifik yang sedang dihitung bobotnya.

d: Dokumen spesifik di mana kata tersebut berada.

$F_{t,d:}$ Jumlah kemunculan istilah *t* dalam dokumen *d*.

Kata (*term*) yang di simbolkan (*t*) merujuk pada satu kata spesifik. Kata spesifik ini terdapat pada sebuah dokumen yang di simbolkan (*d*) yang berguna sebagai wadah untuk kata (*t*) berada. Kemudian pada, rumus (*Jumlah kemunculan kata t*) merupakan proses menghitung manual berapa banyak (*t*) di dalam (*d*). Kemudian, rumus (*Total seluruh kata dalam dokumen d*) merupakan pembanding untuk total jumlah kata (*t*) yang muncul pada suatu dokumen (*d*).

Terdapat beberapa variasi formula yang dapat digunakan untuk menghitung nilai TF, kegunaannya bergantung pada kebutuhan dan karakteristik data untuk dianalisis (Hanani, 2023). Beberapa variasi, diantaranya adalah:

a) *Binary TF*, formula ini hanya mempertimbangkan ada atau tidaknya suatu kata dalam dokumen. Jika kata tersebut ada, nilainya 1 dan jika tidak muncul nilainya 0.
b) *Raw TF*, formula ini dihitung berdasarkan jumlah kemunculan asli sebuah kata dalam dokumen. Contohnya, jika kata muncul 7 kali, maka nilai TF nya 7.
c) *Normalization TF*, formula ini membandingkan kemunculan kata dengan kata yang sering muncul dalam dokumen yang sama. Tujuannya untuk mencegah nilai TF terlalu dipengaruhi oleh panjang dokumen.
d) *Logarithmic TF*, formula ini menggunakan fungsi logaritma (operasi matematika) untuk menilai TF. Tujuannya untuk menghindari penumpukkan kata yang memiliki frekuensi yang sangat tinggi namun tidak relevan.

b. Inverse Document Frequency (IDF)

Inverse Document Frequency (IDF) digunakan untuk mengukur seberapa penting atau unik sebuah kata di dalam semua kumpulan dokumen. IDF digunakan untuk mengurangi dominasi kata-kata yang sering muncul di semua dokumen, contohnya kemunculan kata umum seperti “yang”, “di”, “dan”, “atau” karena kata-kata ini tidak memiliki makna informatif terhadap dokumen (Andayani & Ryansyah, 2017). Dalam penerapannya menggunakan rumus sebagai Berikut:

$$\boldsymbol{IDF(t) = log \frac{N}{df_t}} \quad (2.2)$$

Keterangan:

IDF(t): Nilai Inverse Document Frequency dalam kata t

log: Fungsi perhitungan matematika yang digunakan untuk memperkecil skala nilai agar tidak terlalu besar perbedannya.

N: Total semua dokumen yang ada di dalam koleksi data yang dianalisis.

df_t: Jumlah dokumen yang di dalamnya terdapat kata t (minimal muncul satu kali).

Nilai Inverse Document Frequency yang disimbolkan (*IDF(t)*) merujuk pada skor akhir yang menunjukkan seberapa unik sebuah kata (*t*) di seluruh dokumen. Dalam (*Total jumlah dokumen dalam d*) yang merupakan jumlah keseluruhan komentar yang sedang dianalisis. Kemudian, sebagai pembandingnya, kita hitung (*Total jumlah dokumen di mana kata t*) muncul, yang merupakan proses menghitung di berapa banyak komentar berbeda kata (*t*) tersebut dapat ditemukan. Hasil perbandingan kedua nilai ini kemudian diolah menggunakan fungsi *log* yang berguna untuk menyeimbangkan dan memperkecil skala skornya agar lebih proporsional.

c. Perhitungan Bobot TF-IDF

Dalam metode ini menggabungkan dua konsep untuk perhitungan bobot, yaitu frekuensi kemunculan kata dalam sebuah dokumen dan *inverse* frekuensi dokumen yang mengandung kata tersebut. Rumus dalam penerapan pembobotan TF-IDF adalah:

$$\boldsymbol{W_{d,t} = TF(t,d) \times IDF(t)} \quad (2.3)$$

Keterangan:

$W_{d,t}$: Hasil akhir dari skor bobot TF-IDF dari kata (*t*) di dalam dokumen (*d*).

TF(t,d): Nilai yang menunjukkan kata (*t*) yang sering muncul dalam satu dokumen (*d*).

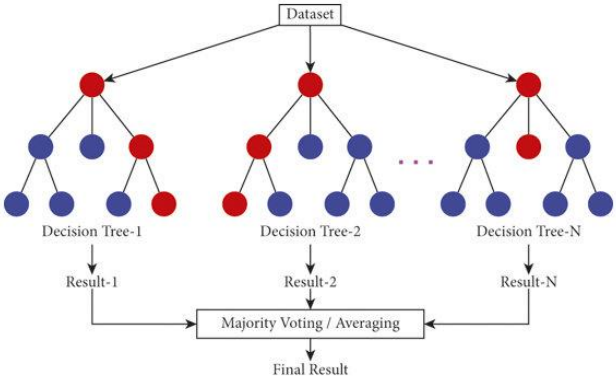
IDF(t): Nilai yang menunjukkan seberapa unik atau langka kata (*t*) di seluruh koleksi dokumen.

Rumus pembobotan TF-IDF yang disimbolkan ($W_{d,t}$) merupakan hasil dari semua proses perhitungan sebelumnya. Nilai ini adalah skor akhir yang kita cari untuk sebuah kata (*t*) dalam sebuah dokumen spesifik (*d*). Untuk mendapatkan skor ini, kita mengalikan dua nilai yang sudah kita hitung: nilai *TF(t,d) yang* menunjukkan seberapa sering kata itu muncul dalam satu komentar, dengan nilai *IDF(t)* yang menunjukkan seberapa istimewa kata itu di antara ribuan komentar lainnya.

2.2.5 Random Forest

Random Forest merupakan salah satu algoritma *machine learning* yang dibangun dengan kumpulan pohon keputusan (*decision tree*). Algoritma ini membentuk sistem klasifikasi berupa kumpulan pohon, di mana setiap pohon dibangun dengan dataset random dan pohon keputusan menentukan hasil prediksinya secara indpenden (Alvanof et al., 2024). Cara kerjanya algoritma *Random Forest* mengumpulkan hasil prediksi dari masing-masing *decision tree* individual dan

melakukan pengumpulan voting berdasarkan suara terbanyak (*majority voting*) (Salman et al., 2024).



Gambar 2.1 Random Forest

(*Sumber: Oleh Trivusi, 2022*)

Berdasarkan gambar 2.1 cara kerja algoritma *Random Forest*, ketika data baru (*instance*) di terapkan ke dalam model, data melalui tahapan analisis oleh semua pohon keputusan (*tree-1*, *tree-2*, *tree-n*) yang ada di dalam *forest*. Setiap pohon keputusan yang dibangun secara independen menggunakan subset data dan fitur yang acak dari data pelatihan. Untuk membangun sebuah pohon keputusan algoritma menentukan pemisahan terbaik (*split*) pada setiap simpul (*node*). Dalam mengukur *split* terbaik, umumnya menggunakan metrik *gini impurity* untuk mencari pemisahan sampel data acak berdasarkan distribusi kelas pada *node*. Nilai pada metrik ini, hanya 0 dan 1, di mana nilai 0 menandakan kemurnian sempurna, sedangkan pada nilai 1 menunjukkan keragaman. Tujuannya adalah untuk memisahkan subset data yang menghasilkan kelompok-kelompok data dengan *gini impurity* rendah (*homogen*) (Arifianto et al., 2022). Rumus gini impurity sebagai berikut:

$$Gini = 1 - \sum_{i=1}^{c} (pi)^2 \tag{2.4}$$

Keterangan:

c: Jumlah total kelas/label

pi: Probabilitas sampel yang termasuk pada kelas i pada suatu node.

Setelah setiap pohon memberikan hasil klasifikasinya, contohnya pada tree-1 hasil klasifikasi sebagai kelas A, sedangkan tree-2 dan tree-3 sebagai kelas B. Kemudian semua hasilnya dikumpulkan dan dilakukan tahap voting berdasarkan hasil atau keputusan terbanyak dari masing-masing pohon keputusan (decision tree) yang disebut majority voting (pengambilan suara mayoritas). Kelas yang paling banyak dipilih oleh mayoritas pohon keputusan akan menjadi prediksi akhir model.

Algoritma ini termasuk dalam kategori *ensemble learning*. *Ensemble learning* merujuk pada penggunaan beberapa model (*classifer*) atau sistem pembelajaran yang digabungkan untuk membuat keputusan atau prediksi yang lebih akurat dibandingkan hanya dengan menggunakan satu model saja (Nugroho, 2025). Keunikan dan kekuatan algoritma ini berasal pada dua metode kerandoman yang diterapkan selama proses pelatihan:

a. *Bagging* (*Bootstrap Agregating*), di mana prosesnya membuat banyak subset data secara acak dari data pelatihan melalui proses *sampling with replacement* (pengambilan sampel dengan pergantian). Maksudnya, setiap data yang terpilih lebih dari satu kali dalam subset, sementara dari data lain tidak terpilih sama sekali. Proses ini memastikan persepektif berbeda pada setiap data (Nugroho, 2025).
b. *Random Feature Selection* adalah proses yang membedakan *random forest* dengan metode *bagging* biasa. Proses ini memungkinkan algoritma mencari fitur terbaik dalam subset fitur acak yang dipilih untuk melakukan pembagian (*split*) saat membangun pohon-pohon.

Algoritma *random forest* memiliki beberapa parameter penting yang berperan dalam pembentukan pohon keputusan acak. Beberapa parameter utama yang paling berpengaruh terhadap hasil prediksi dan mengurangi resiko *overfitting* (Nugroho, 2025).

1. *n_estimators*: Parameter ini menentukan jumlah pohon keputusan (*decision tree*) yang dibangun dalam hutan (*forest*).
2. *max_features*: Parameter ini menghitung jumlah fitur yang dipilih secara acak pada setiap *node* (simpul) untuk pembagian data terbaik.
3. *min_samples_leaf*: Parameter ini berfungsi untuk menentukan jumlah minimum sampel yang harus ada pada sebuah *leaf node* (daun terminal).
4. *min_samples_split*: Parameter ini berfungsi mentapkan jumlah minimum sampel yang dibutuhkan pada *node* agar *node* bisa dibagi lagi.
5. *max_depth*: Parameter ini membatasi kedalaman atau level maksimum dari setiap pohon keputusan.

2.2.6 Malaka Project

Nama “Malaka” terinspirasi dari semangat perjuangan Tan Malaka, seorang tokoh revolusioner Indonesia yang dikenal atas dedikasinya dalam memperjuangkan pendidikan dan kebebasan berpikir. Namun, Malaka Project tidak sepenuhnya mengadopsi ideologi Tan Malaka, melainkan hanya mengambil semangatnya dalam memperjuangkan pendidikan dan menumbuhkan kesadaran berpikir kritis di tengah masyarakat. Malaka Project dibangun oleh sembilan *founder* dengan latar belakang yang berbeda-beda, namun memiliki tujuan yang sama dalam meningkatkan literasi khususnya generasi muda melalui konsumsi konten bermanfaat. Para *founder* yang menjadi representasi dari Malaka Project adalah Ferry Irwandi, Cania Citta,

Jerome Polin, Fathia Izzati, Aurelia Vizal, Angellie Nabila, Dea Anugrah, Coki Pardede, dan Rizky Adi Prakoso.

2.2.7 Confusion Matrix

Confusion Matrix adalah tabel matriks, yang berguna untuk mengukur performa klasifikasi pada model *machine learning*, parameternya berdasarkan *accuracy*, *precision*, *recall*, dan *f1-score*. Di mana sebuah model dikatakan baik jika nilai performanya tinggi yang diukur melalui metode ini, biasanya dilakukan dalam pengukuran polarisasi opini (Fauzan & Hikmah, 2022). Pada umumnya hasil klasifikasi memiliki perbandingan nilai aktual dan nilai prediksi, sehingga memudahkan dalam mengukur tingkat akurasi dan kesalahan prediksi (Amin, 2023).

		Prediksi	
		Positif	Negatif
Aktual	Positif	TP	FN
	Negatif	FP	TN

Gambar 2.2 Confusion Matrix

Sumber: Peneliti, 2025

Terdapat empat komponen penting dalam membentuk matriks evaluasi, antara lain:

a. TP (True Positive)

Kondisi ini terjadi berdasarkan prediksi model mendapatkan hasil positif, sesuai dengan label aslinya (komentar positif).

b. TN (True Negative)

Sama seperti TP (*True Positive*), kondisi ini terjadi berdasarkan prediksi model mendapatkan hasil negatif, sesuai dengan label aslinya (komentar negatif).

c. FP (False Positive)

Berbanding terbalik dengan hasil prediksi sebelumnya atau prediksi yang benar (*true*), kondisi ini terjadi ketika model salah

memprediksi kelas positif, prediksi tersebut tidak sesuai dengan labelnya (komentar positif).

d. FN (False Negative)

Kondisi ini juga serupa dengan FP (*False Positive*), ketika model salah memprediksi kelas negatif, kondisi yang berlawanan dengan label (komentar negatif).

Confusion Matrix memiliki pengukuran kinerja yang lebih kompleks, selain dari ke-empat komponen diatas. Adapaun *performance measurements* (pengukuran kinerja) yaitu *accuracy*, *precision*, *recall*, dan *f1-score* (Sathyanarayanan, 2024).

1. Accuracy (Akurasi)

Accuracy merupakan matriks yang umum digunakan dalam mengukur presentase prediksi yang benar dari keseluruhan hasil prediksi yang dibuat oleh model *machine learning*. Rumus untuk menghitung *accuracy* sebagai berikut:

$$\boldsymbol{Accuracy} = \frac{\boldsymbol{TP + TN}}{\boldsymbol{TP + FP + TN + FN}} \quad (2.5)$$

Keterangan:

TP (True Positive): Jumlah data positif yang diprediksi benar sebagai positif.

TN (True Negative): Jumlah data negatif yang diprediksi benar sebagai negatif.

FP (False Positive): Jumlah data positif yang salah diprediksi sebagai positif.

FN (False Negative): Jumlah data negatif yang salah diprediksi sebagai negatif.

Rumus *accuracy* (akurasi) digunakan untuk menghitung rasio jumlah prediksi dengan total seluruh data uji. Pada rumus **(TP+TN)** merupakan jumlah keseluruhan prediksi benar. **TP** data positif yang ditebak benar oleh model, sementara **TN** data

negatif yang ditebak benar. Pada rumus **(TP+FP+TN+FN)** merupakan jumlah keseluruhan data atau hasil dari prediksi yang dilakukan model, bersifat *general* tidak terdefinisi baik atau benar.

2. Precision (Presisi)

Precision merupakan matriks yang digunakan untuk mengukur tingkat ketepatan dari prediksi positif yang dihasilkan oleh model. Matriks ini menghitung jumlah data yang sebenarnya positif dari total semua data yang memiliki nilai positif. Rumus untuk menghitung *precision* sebagai berikut:

$$\boldsymbol{Precision} = \frac{\boldsymbol{TP}}{\boldsymbol{TP} + \boldsymbol{FP}} \tag{2.6}$$

Keterangan:

TP (True Positive): Jumlah data positif yang diprediksi benar sebagai positif.

FP (False Positive): Jumlah data positif yang salah diprediksi sebagai positif.

Rumus *precision* (presisi) berfokus pada kualitas produksi data positif yang dihasilkan model. Pada rumus **TP** menggambarkan jumlah data prediksi benar positif. Pada rumus **(TP+FP)** merupakan total semua data yang diprediksi sebagai positif oleh model, terlepas dari hasilnya benar positif atau salah prediksi.

3. Recall (Sensitivitas)

Recall biasa dikenal *True Positive Rate* (TPR), digunakan untuk mengukur kemampuan model dalam mengidentifikasi semua data yang sebenarnya positif. Matriks ini mengukur seluruh data positif yang berhasil diprediksi oleh model dengan benar. Rumus untuk menghitung *recall* sebagai berikut:

$$\boldsymbol{Recall} = \frac{\boldsymbol{TP}}{\boldsymbol{TP} + \boldsymbol{FN}} \tag{2.7}$$

Keterangan:

TP (True Positive): Jumlah data positif yang diprediksi benar sebagai positif.

FN (False Negative): Jumlah data negatif yang salah diprediksi sebagai negatif.

Rumus *recall* (sensitivitas) mengukur kemampuan model dalam menemukan kembali semua data yang sebenarnya positif. Pada rumus **TP** merepresentasikan jumlah prediksi data yang benar. Pada rumus **(TP+FN)** merepresentasikan keseluruhan data yang sebenarnya positif di dalam dataset, baik yang berhasil ditemukan kembali oleh model maupun yang terlewat.

4. F1-Score

F1-Score merupakan matriks evaluasi yang menggabungkan tahapan matriks *precision* dan *recall* menggunakan rata-rata harmonic (*harmonic mean*). Matriks ini akan menggabungkan hasil dari keduanya, jika salah satunya memiliki nilai yang sangat rendah maka hasil dari *f1-score* juga rendah. Rumus untuk menghitung *f1-score* sebagai berikut:

$$\boldsymbol{F1 - Score = \frac{2 \times Precision \times Recall}{Precision + Recall}} \quad (2.8)$$

Keterangan:

Precision: Nilai presisi yang diperoleh dari perhitungan sebelumnya. Nilai ini merepresentasikan seberapa akurat prediksi positif yang dilakukan oleh model.

Recall: Nilai *recall* atau sensitivitas yang diperoleh dari perhitungan sebelumnya. Nilai ini mengukur kemampuan model untuk mengidentifikasi data yang sebenarnya positif.

Rumus *f1-score* merupakan gabungan matriks dari *precision* dan *recall*. Rumus ini dihitung berdasarkan *harmonic mean*

terhadap hasil keduanya dan menjadikan satu matriks tunggal. *Harmonic mean* sangat mempengaruhi hasil matriks *f1-score* jika salah satu diantara nilai *precision* atau *recall* mendapatkan skor sangat rendah

2.2.8 Verifikasi Informasi (Tabayyun)

Pada era informasi yang mudah didapat secara cepat, perlu diperhatikan pentingnya akhlak dalam menyebarkan informasi kepada masyarakat luas agar tidak terjadi kesalahpahaman antar sesama. Menurut pandangan islam seseorang harus memiliki akhlak yang baik dalam bertindak terutama dan mampu bertanggung jawab atas tindakannya. Dalam menyebarkan sebuah informasi, istilah akhlak ini disebut *tabbayun* di mana berkaitan dengan proses verifikasi, klarifikasi, dan memvalidasi kebenaran terhadap informasi yang diterima sebelum menjadikannya sebagai dasar untuk bersikap, berpendapat, mengambil keputusan, atau bertindak (Azzuhri, 2020). *Tabayyun* menjadi sangat penting untuk diterapkan pada era digital, di tengah banyak masalah yang disebabkan oleh penyebaran informasi yang keliru (*hoax*). Sikap yang selalu mempercayai tanpa melakukan verifikasi beresiko menimbulkan keputusan yang salah dan merugikan diri sendiri serta orang lain.

Landasan utama perintah menerapkan *tabayyun* tertuang dalam Q.S. Al-Hujurat ayat 6 yang berbunyi:

يٰٓاَيُّهَا الَّذِيْنَ اٰمَنُوْٓا اِنْ جَاۤءَكُمْ فَاسِقٌۢ بِنَبَاٍ فَتَبَيَّنُوْٓا اَنْ تُصِيْبُوْا قَوْمًاۢ بِجَهَالَةٍ فَتُصْبِحُوْا عَلٰى مَا فَعَلْتُمْ نٰدِمِيْنَ ﴿٦﴾

Artinya:

"*Wahai orang-orang yang beriman, jika seorang fasik datang kepadamu membawa seuatu berita, maka telitilah kebenarannya agar kamu tidak mencelakakan suatu kaum karena kebodohan* (*kecerobohan*) *yang akhirnya kamu menyesali perbuatanmu itu"*.

Secara bahasa *tabayyun* berasal dari kata "*Baana*" artinya sesuatu yang jelas, sedangkan kalimat "*fattabayanu*" dalam ayat surah Al-Hujurat ayat 6 merupakan bentuk anjuran yang memiliki makna, anjuran dalam melakukan validitas atau memeriksa kebenarannya, melalui proses investigasi atau klarifikasi untuk membuktikan sesuatu yang belum jelas atau samar menjadi terpercaya (Azzuhri, 2020). Dalam sejarahnya turunnya ayat ini berkaitan dengan peristiwa ketika Rasulullah SAW mengutus Al-Walid bin 'Ubqah untuk mengambil zakat dari kaum Bani Musthaliq. Karena memiliki latar belakang sejarah yang kurang baik Al-Walid memutuskan kembali sebelum sampai dan melaporkan kepada Rasulullah SAW bahwa Bani Mushtaliq telah murtad dan menolak membayar zakat. Namun, perwakilan dari Bani Mushtaliq datang dan menyatakan kaum nya setia kepada islam, sehingga perpecahan yang disebabkan kekeliruan informasi dapat dihindari. Peristiwa ini memiliki pelajaran bahwa setiap informasi yang diterima dari orang fasik tanpa melakukan pengecekan dapat menyebabkan musibah besar, melakukan tuduhan kepada orang lain atas dasar kebodohan, dan berujung pada penyesalan mendalam (Qur'an Kemenag, 2022).

2.2.9 Fikih Perkataan

Dalam agama islam menjaga lisan merupakan sebuah keharusan bagi orang-orang beriman, karena dengan menjaga perkataan kita bisa menjaga keharmonisan sesama di lingkungan masyarakat. Hal tersebut dijelaskan dalam kitab suci Al-Qur'an, di mana topik komunikasi banyak ditemukan lewat istilah *al-qaul* (قول) mengandung makna sesuatu yang terucap dari lisan secara sadar dan menggambarkan ekspresi. Terdapat enam macam *qaul* yang dijadikan panduan, yaitu *qaulan sadida*, *qaulan ma'rufa*, *qaulan baligha*, *qaulan maysura*, *qaulan layyina*, dan *qaulan karima* (Nurhasanah & Suherman, 2022).

Pertama *qaulan sadida* memiliki makna perkataan yang tulus dan jujur. Artinya perkataan disampaikan berdasarkan fakta dan niat baik, bukan untuk menipu. Kata ini terkandung di dalam Q.S. An-Nisa ayat 9 yang berbunyi:

وَلْيَخْشَ الَّذِيْنَ لَوْ تَرَكُوْا مِنْ خَلْفِهِمْ ذُرِّيَّةً ضِعٰفًا خَافُوْا عَلَيْهِمْۖ فَلْيَتَّقُوا اللّٰهَ وَلْيَقُوْلُوْا قَوْلًا سَدِيْدًا ۝٩

Artinya: "*Hendaklah merasa takut orang-orang yang seandainya (mati) meninggalkan setelah mereka, keturunan yang lemah (yang) mereka khawatir terhadapnya. Maka, bertakwalah kepada Allah dan berbicalah dengan tutur kata yang benar (dalam hal menjaga hak-hak keturunannya)*".

Kedua *qaulan ma'rufa* memiliki makna perkataan yang baik, sopan, dan tepat pada situasinya. Artinya seseorang harus memahami kondisi pendengar dan menyampaikan pesan dengan sopan yang sesuai dengan syariat islam. Kata ini terkandung dalam Q.S. Al-Ahzab ayat 32 yang berbunyi:

يٰنِسَاۤءَ النَّبِيِّ لَسْتُنَّ كَاَحَدٍ مِّنَ النِّسَاۤءِ اِنِ اتَّقَيْتُنَّ فَلَا تَخْضَعْنَ بِالْقَوْلِ فَيَطْمَعَ الَّذِيْ فِيْ قَلْبِهٖ مَرَضٌ وَّقُلْنَ قَوْلًا مَّعْرُوْفًاۚ ۝٣٢

Artinya: "*Wahai istri-istri Nabi, kamu tidaklah seperti perempuan-perempuan yang lain jika kamu bertakwa. Maka, janganlah kamu merendahkan suara (dengan lemah lembut yang dibuat-buat) sehingga bangkit nafsu orang yang ada penyakit dalam hatinya dan ucapkanlah perkataan yang baik*".

Ketiga *qaulan baligha* memiliki makna penyampaian yang efeketif, jelas, dan tidak bertele-tele. Artinya perkataan langsung ditujukan kepada pokok permasalahan dan membekas di jiwa pendengarnya dengan kesan yang baik. Kata ini terkandung dalam Q.S. An-Nisa ayat 63:

اُولٰۤىِٕكَ الَّذِيْنَ يَعْلَمُ اللّٰهُ مَا فِيْ قُلُوْبِهِمْ فَاَعْرِضْ عَنْهُمْ وَعِظْهُمْ وَقُلْ لَّهُمْ فِيْٓ اَنْفُسِهِمْ قَوْلًا ۢ بَلِيْغًا ۝٦٣

Artinya: "*Mereka itulah orang-orang yang Allah ketahui apa yang ada di dalam hatinya. Oleh karena itu, berpalinglah dari mereka,*

nasihatilah mereka, dan katakanlah kepada mereka perkataan yang membekas pada jiwanya".

Keempat *qaulan maysura* memiliki makna ucapan yang mudah dipahami, artinya perkataan yang diucapkan memberikan kesan menenangkan dan tidak memberatkan lawan bicara. Kata ini terkandung dalam Q.S. Al-Isra ayat 28 yang berbunyi:

وَاِمَّا تُعْرِضَنَّ عَنْهُمُ ابْتِغَاۤءَ رَحْمَةٍ مِّنْ رَّبِّكَ تَرْجُوْهَا فَقُلْ لَّهُمْ قَوْلًا مَّيْسُوْرًا ﴿٢٨﴾

Artinya: "*Jika (tidak mampu membantu sehingga) engkau (terpaksa) berpaling dari mereka untuk memperoleh rahmat dari Tuhanmu yang engkau harapkan, ucapkanlah kepada mereka perkataan yang lemah lembut*".

Kelima *qaulan layyina* memiliki makna ucapan lembut, halus, atau tidak keras, artinya setiap ucapan baik itu berupa kritik atau saran disampaikan dengan nada yang menenangkan dan menghargai. Kata ini terkandung dalam Q.S. Thaha ayat 44 yang berbunyi:

فَقُوْلَا لَهٗ قَوْلًا لَّيِّنًا لَّعَلَّهٗ يَتَذَكَّرُ اَوْ يَخْشٰى ﴿٤٤﴾

Artinya: "*Berbicaralah kamu berdua kepadanya (Fir'aun) dengan perkataan yang lemah lembut, mudah-mudahan dia sadar atau takut*".

Keenam *qaulan karima* memiliki makna mulia atau terhomat, artinya menyampaikan pesan dengan sopan kepada pihak yang wajib dihormati seperti orang tua, guru, dan orang yang lebih tua. Kata ini terkandung dalam Q.S. Al-Isra' ayat 23 yang berbunyi:

وَقَضٰى رَبُّكَ اَلَّا تَعْبُدُوْٓا اِلَّآ اِيَّاهُ وَبِالْوَالِدَيْنِ اِحْسَانًاۗ اِمَّا يَبْلُغَنَّ عِنْدَكَ الْكِبَرَ اَحَدُهُمَآ اَوْ كِلٰهُمَا فَلَا تَقُلْ لَّهُمَآ اُفٍّ وَّلَا تَنْهَرْهُمَا وَقُلْ لَّهُمَا قَوْلًا كَرِيْمًا ﴿٢٣﴾

Artinya: "*Tuhanmu telah memerintahkan agar kamu jangan menyembah selain Dia dan hendaklah berbuat baik kepada ibu bapak. Jika salah seorang di antara keduanya atau kedua-duanya sampai berusia lanjut dalam pemeliharaanmu, maka sekali-kali janganlah engkau mengatakan kepada keduanya perkataan "ah" dan janganlah engkau membentak keduanya, serta ucapkanlah kepada keduanya perkataan yang baik*".

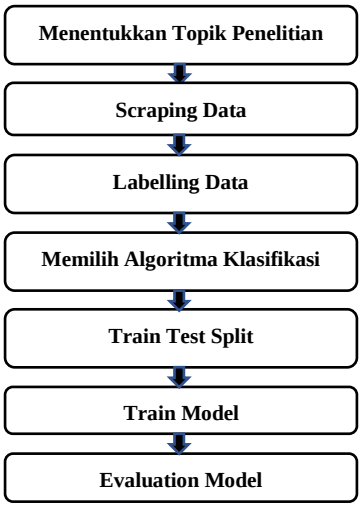
BAB III
METODE PENELITIAN

3.1 Jenis Penelitian

Penelitian ini menggunakan pendekatan kuantitatif dengan menerapkan metode *Text Mining* yang bertujuan untuk mengolah dan mengklasifikasikan data. Fokus utama penelitian ini mengklasifikasikan sentimen ke dalam tiga kelas (positif, negatif, netral). Proses klasifikasi dilakukan secara oleh algoritma klasifikasi *Random Forest.* Hasil dari prediksi akan dilakukan evaluasi berdasarkan metrik *accuracy*, *precision*, *recall*, *f1-score* yang perhitungannya didasarkan pada tabel *Confusion Matrix*.

3.2 Alur Penelitian

Beberapa tahapan yang dilakukan dalam penelitian ini meliputi proses yang sistematis dan tertstruktur bertujuan menjamin validitas setiap prosesnya. Seperti menentukan topik penelitian, melakukan pengambilan data atau *Scraping*, melabeli data atau proses *labelling* secara manual ke dalam tigas kelas sentimen, memilih algoritma klasifikasi, melakukan pembagian terhadap data pelatihan dan data pengujian, melatih model algoritma untuk klasifikasi, dan terakhir mengevaluasi hasil klasifikasi. Tahapan ini disusun untuk menggambarkan proses transformasi data tidak terstruktur menjadi informasi yang memiliki makna melalui pendekatan analisis sentimen. Melalui rangkaian tahapan ini, peneliti berupaya membentuk model klasifikasi yang tidak hanya berfungsi secara teknis, tetapi juga memiliki cara kerja yang jelas dalam menggambarkan respon masyarakat terhadap topik korupsi. Secara garis besar, alur ini di bentuk untuk memastikan algoritma yang digunakan dalam mengelola ribuan fitur kata menjadi lebih efektif.

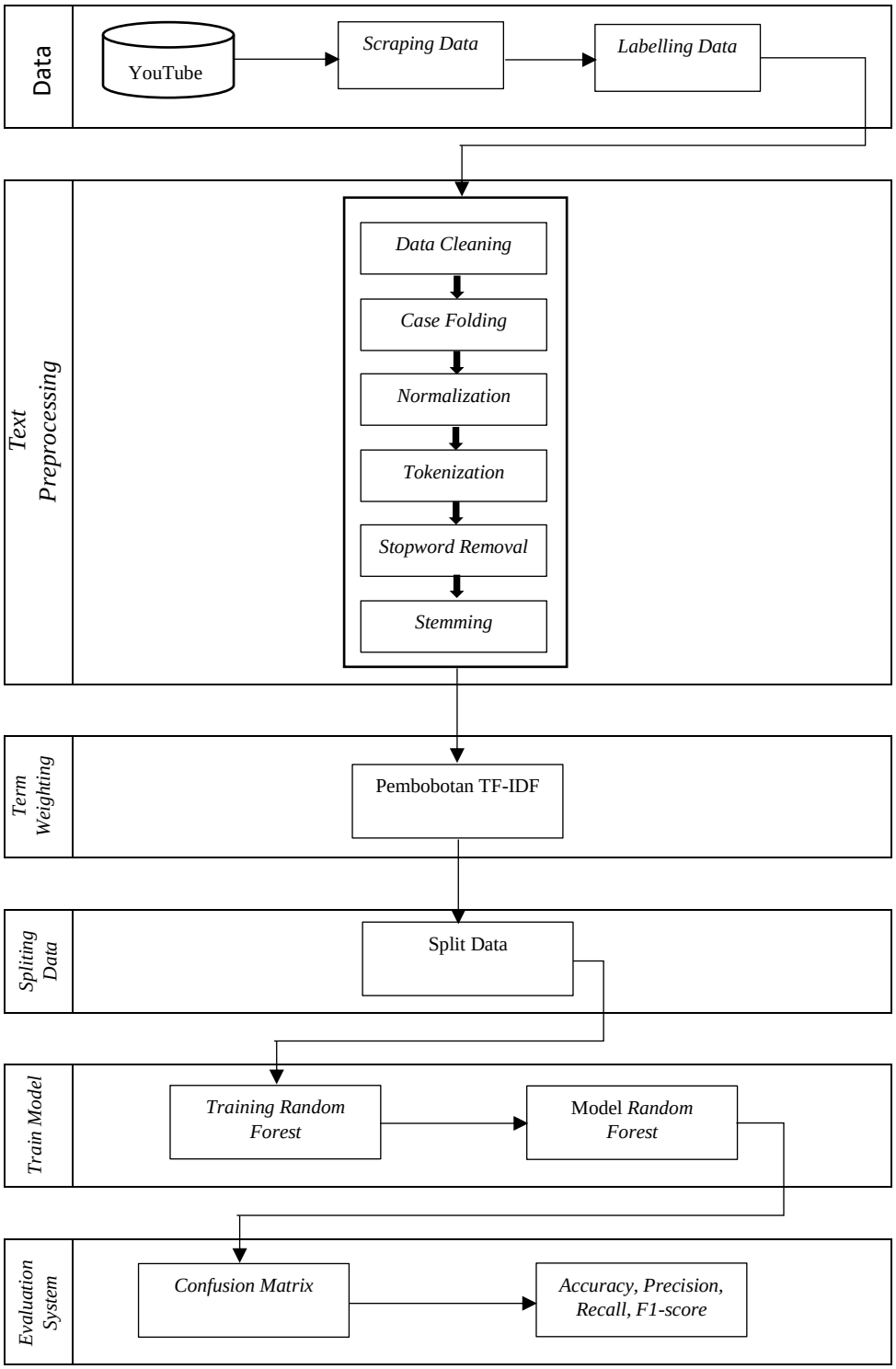


Gambar 3.1 Alur Penelitian

(*Sumber: Hasil Olah Data*)

Berdasarkan Gambar 3.1, alur penelitian ini dirancang secara sistematis melalui tujuh tahapan utama yang saling berkaitan. Pertama, menentukan topik penelitian dilakukan berdasarkan banyaknya kasus korupsi di Indonesia serta relevansinya dengan konten YouTube Malaka Project. Kedua, melakukan *Scraping data* yaitu pengumpulan data komentar dengan aplikasi *Google App Script* dengan memanfaatkan akses *API* resmi. Ketiga, *Labelling Data* di mana seluruh komentar diberi label sebagai validasi awal yang dikategorikan menjadi, label positif, label netral, dan label negatif. Keempat, memilih algoritma klasifikasi menggunakan *Random Forest Classifier* yang dinilai sangat kuat dalam menangani data teks berjumlah besar dan tidak terstruktur. Kelima, *Train-Test Split* yaitu pembagian dataset menjadi dua skenario berbeda, yakni 80:20 dan 70:30. Keenam, *Train Model* merupakan proses pelatihan algoritma menggunakan data latih yang telah melewati tahap *preprocessing* dan ekstraksi fitur *TF-IDF*. Terakhir, *Evaluation Model* dilakukan menggunakan tabel *Confusion Matrix* untuk mengukur nilai *accuracy*, *precision*, *recall*, dan *f1-score* agar hasil dari prediksi model dapat dipertanggungjawabkan secara ilmiah.

3.3 Desain Sistem



Gambar 3.2 Desain Sistem

(Sumber: *Hasil Olah Data*)

Pada gambar 3.2 adalah alur desain sistem, dimulai dari pengambilan data komentar YouTube pada kanal Malaka Project, kemudian dilakukan pelabelan manual. Data yang sudah dilabeli melalui tahap *preprocessing* yang mencakup *data cleaning*, *case folding*, *tokenizing*, *normalization*, *stopword removal*, dan *stemming*. Selanjutnya dilakukan pembobotan kata menggunakan metode TF-IDF (*Term Frequency – Inverse Document Frequency*). Selanjutnya pembagian data menjadi data latih dan data uji dengan rasio 70:30 dan 80:20. Pelatihan model klasifikasi *Random Forest* menggunakan data latih dan dilakukan pengujian pada data uji sebagai hasil prediksi. Terakhir, hasil prediksi model akan di evaluasi dengan tabel matriks *Confusion Matrix* berdasarkan metrik *accuracy*, *precision*, *recall*, dan *f1-score*.

3.3.1 Scraping Data

Scraping Data merupakan proses mengolah data, konteks disini data komentar atau mengekstraksi data dari situs web untuk menemukan pola tersembunyi (Yuniar et al., 2022). Data yang di ekstraksi dari salah satu *platform* media sosial *YouTube* pada *kanal* Malaka Project, dengan berfokus pada konten membahas korupsi. Pengambilan data dilakukan pada tanggal 20 Oktober 2025 dan data disimpan dalam format CSV (*Comma Separated Values*).

3.3.2 Labelling

Pada tahapan ini pemberian label dilakukan secara manual yang dilakukan oleh tiga ahli bahasa, pelabelan ini diklasifikasikan ke dalam tiga kategori, yaitu label positif, netral, dan negatif. Tujuan pemberian label untuk memudahkan model dalam memproses data karena terdapat informasi tambahan atau kelas dalam data (Kamil et al., 2025). Sebagai contoh pelabelan manual pada data komentar:

Tabel 3.1 Contoh Pelabelan Data

No	Teks	Label
1	Gila banget orang2 Pertamina yg kemaruk duit ini!	Positif
2	Coba feri tanya ahok kenapa dia membiarkan korupsi di pertamina	Netral
3	Yg paling jahat lagi korupsi haji bang	Netral
4	Setiap hari Terima kabar buruk dari negara	Positif
5	Dengar ini jadi tambah emosi	Positif
6	Catet…..hakimnya…...hujat dan 39error…..selidiki sutradaranya…..cari dan temukan!!!!	Negatif
7	Ekonomi kapitalis biasanya ga disukai komunis Ðÿ˜…	Negatif
8	Kualitas penegak hukum kita buruk sekali ya	Positif
9	Justice is dead	Positif
10	Di indo jadi hakim asal mau nurut sama penguasa	Positif

3.3.3 Preprocessing Text

Preprocessing merupakan tahapan pembersihan pada data mentah yang dilakukan sebelum data dilatih oleh model. Tahapan ini merupakan pengelolaan data yang bertujuan mengubah data tidak terstruktur menjadi data tersutruktur dan dapat diolah oleh algoritma (Jamil et al., 2024). Proses ini sangat penting dilakukan untuk memastikan kualitas data karena hasil klasifikasi sangat bergantung pada kebersihan dan konsistensi data yang digunakan. Pada penelitian ini tahapannya meliputi *cleaning*, *case folding*, *tokenizing*, *normalization*, *stopword removal*, dan *stemming*.

a. *Cleaning*

Proses ini meliputi pembersihan pada data, pembersihan yang dilakukan pada elemen atau symbol yang tidak relevan terhadap teks, seperti emoji, hashtag, URL, dan lainnya. Tujuannya untuk menghasilkan teks yang lebih bersih dan terstruktur, agar mudah diproses oleh tahapan selanjutnya. Penerapan proses *cleaning* pada data mentah seperti gambar berikut:

Tabel 3.2 Contoh Tahapan Cleaing

Data Mentah	Cleaning
Gila banget orang2 Pertamina yg kemaruk duit ini!	Gila banget orang Pertamina yg kemaruk duit ini
Coba feri tanya ahok kenapa dia membiarkan korupsi di pertamina	Coba feri tanya ahok kenapa dia membiarkan korupsi di pertamina
Yg paling jahat lagi korupsi haji bang	Yg paling jahat lagi korupsi haji bang
Setiap hari Terima kabar buruk dari negara	Setiap hari Terima kabar buruk dari negara
Dengar ini jadi tambah emosi	Dengar ini jadi tambah emosi
Catet.....hakimnya.....hujat dan teror.....selidiki sutradaranya.....cari dan temukan!!!!	Catethakimnyahujat dan terorselidiki sutradaranyacari dan temukan
Ekonomi kapitalis biasanya ga disukai komunis ðŸ˜…	Ekonomi kapitalis biasanya ga disukai komunis
Kualitas penegak hukum kita buruk sekali ya	Kualitas penegak hukum kita buruk sekali ya
Justice is dead	Justice is dead
Di indo jadi hakim asal mau nurut sama penguasa	Di indo jadi hakim asal mau nurut sama penguasa

b. *Case Folding*

Case folding merupakan proses menyelaraskan teks menjadi huruf kecil atau *lowercase* (Yusrana et al., 2024). Tahapan ini dilakukan sebagai upaya penyeragaman bentuk huruf agar tidak terjadi perbedaan makna ketika dibaca oleh sistem. Tujuannya agar teks lebih mudah di analisis oleh algoritma, Dengan, bentuk huruf yang lebih konsisten dan jumlah varias kata berkurang, proses ini juga sangat membantu tahapan-tahapan selanjutnya dan meningkatkan kualitas pembobotan kata. Berikut merupakan penerapan *case folding* dalam data teks:

Tabel 3.3 Contoh Tahapan Case Folding

Cleaning	Case Folding
Gila banget orang Pertamina yg kemaruk duit ini	gila banget orang pertamina yg kemaruk duit ini
Coba feri tanya ahok kenapa dia membiarkan korupsi di pertamina	coba feri tanya ahok kenapa dia membiarkan korupsi di pertamina
Yg paling jahat lagi korupsi haji bang	yg paling jahat lagi korupsi haji bang
Setiap hari Terima kabar buruk dari negara	setiap hari terima kabar buruk dari negara
Dengar ini jadi tambah emosi	dengar ini jadi tambah emosi
Catethakimnyahujat dan terorselidiki sutradaranyacari dan temukan	catethakimnyahujat dan terorselidiki sutradaranyacari dan temukan
Ekonomi kapitalis biasanya ga disukai komunis	ekonomi kapitalis biasanya ga disukai komunis
Kualitas penegak hukum kita buruk sekali ya	kualitas penegak hukum kita buruk sekali ya
Justice is dead	justice is dead
Di indo jadi hakim asal mau nurut sama penguasa	di indo jadi hakim asal mau nurut sama penguasa

c. *Normalization*

Normalization merupakan tahapan merubah kata yang tidak baku (tidak sesuai kaidah penulisan bahasa Indonesia) menjadi kata baku seperti pada kamus KBBI (Kamus Besar Bahasa Indonesia). Perubahan yang dilakukan agar kata-kata yang memiliki makna sama dapat dikelompokkan ke dalam kata yang seragam oleh sistem. Tujuannya agar teks lebih mudah dipahami oleh algoritma dalam mengolah data klasifikasi. Selain itu, tahapan ini juga membantu mengurangi variasi kata sehingga data bisa lebih konsisten. Berikut contoh proses *normalization* pada data teks:

Tabel 3.4 Contoh Tahapan Normalization

Case Folding	Normalization
gila banget orang pertamina yg kemaruk duit ini	gila banget orang pertamina yang rakus uang ini
coba feri tanya ahok kenapa dia membiarkan korupsi di pertamina	coba ferry tanya ahok kenapa dia membiarkan korupsi di pertamina
yg paling jahat lagi korupsi haji bang	yang paling jahat lagi korupsi haji bang
setiap hari terima kabar buruk dari negara	setiap hari terima kabar buruk dari negara
dengar ini jadi tambah emosi	dengar ini jadi tambah emosi
catethakimnyahujat dan terorselidiki sutradaranyacari dan temukan	catet hakim nya hujat dan terror selidiki sutradaranya cari dan temukan
ekonomi kapitalis biasanya ga disukai komunis	ekonomi kapitalis biasanya tidak disukai komunis
kualitas penegak hukum kita buruk sekali ya	kualitas penegak hukum kita buruk sekali iya
justice is dead	hukum telah mati
di indo jadi hakim asal mau nurut sama penguasa	di indonesia jadi hakim asal mau nurut sama penguasa

d. Tokenizing

Tokenizing merupakan proses pengelompokkan teks menjadi segmen-segmen yang lebih kecil dan terstruktur (Muzaki et al., 2024). Segmen atau token berisi kata, frasa, atau unit teks yang terpisah dari kalimat aslinya. Tujuannya agar sistem lebih mudah menganalisis klasifikasinya. Melalui tahapan ini, setiap kata menjadi fitur independent sebelum masuk ke tahap selanjutnya. Berikut proses penerapan *tokenizing* pada data teks:

Tabel 3.5 Contoh Tahapan Tokenizing

Normalization	Tokenizing
gila banget orang pertamina yang rakus uang ini	[gila, banget, orang, pertamina, yang, rakus, uang, ini]
coba ferry tanya ahok kenapa dia membiarkan korupsi di pertamina	[coba, ferry, tanya, ahok, kenapa, dia, membiarkan, korupsi, di, pertamina]
yang paling jahat lagi korupsi haji bang	[yang, paling, jahat, lagi, korupsi, haji, bang]
setiap hari terima kabar buruk dari negara	[setiap, hari, terima, kabar, buruk, dari, negara]
dengar ini jadi tambah emosi	[dengar, ini, jadi, tambah, emosi]
catet hakim nya hujat dan terror selidiki sutradaranya cari dan temukan	[catet, hakimnya, hujat, dan, terror, selidiki, sutradaranya, cari, dan, temukan]
ekonomi kapitalis biasanya tidak disukai komunis	[ekonomi, kapitalis, biasanya, tidak, disukai, komunis]
kualitas penegak hukum kita buruk sekali iya	[kualitas, penegak, hukum, kita, buruk, sekali, iya]
hukum telah mati	[hukum, telah, mati]
di indonesia jadi hakim asal mau nurut sama penguasa	[di, indonesia, jadi, hakim, asal, mau, nurut, sama, penguasa]

e. *Stopword Removal*

Tokenizing merupakan proses penghapusan imbuhan kata atau kata hubung pada teks yang tidak memiliki makna (Jamil et al., 2024). Pada penelitian ini proses *stopword* menggunakan Sastrawi untuk menghilangkan kata yang sudah terdapat di daftar kata *stopword* berbahasa Indonesia (Pratama & Rivan, 2024). Dengan tahapan ini teks menjadi lebih ringkas dan bersih sebelum lanjut ke tahap berikutnya. Contoh penerapan *stopword removal* pada data teks:

Tabel 3.6 Contoh Tahapan Stopwords Removal

Tokenizing	Stopword Removal
[gila, banget, orang, pertamina, yang, rakus, uang, ini]	[gila, banget, orang, pertamina, rakus, uang, ini]
[coba, ferry, tanya, ahok, kenapa, dia, membiarkan, korupsi, di, pertamina]	[coba, ferry, tanya, ahok, membiarkan, korupsi, pertamina]
[yang, paling, jahat, lagi, korupsi, haji, bang]	[paling, jahat, korupsi, haji, bang]
[setiap, hari, terima, kabar, buruk, dari, negara]	[hari, terima, kabar, buruk, negara]
[dengar, ini, jadi, tambah, emosi]	[dengar, jadi, tambah, emosi]
[catet, hakimnya, hujat, dan, terror, selidiki, sutradaranya, cari, dan, temukan]	[catet, hakimnya, hujat, terror, selidiki, sutradaranya, cari, temukan]
[ekonomi, kapitalis, biasanya, tidak, disukai, komunis]	[ekonomi, kapitalis, biasanya, disukai, komunis]
[kualitas, penegak, hukum, kita, buruk, sekali, iya]	[kualitas, penegak, hukum, buruk, sekali, iya]
[hukum, telah, mati]	[keadilan, mati]
[di, indonesia, jadi, hakim, asal, mau, nurut, sama, penguasa]	[indonesia, jadi, hakim, asal, mau, nurut, sama, penguasa]

f. *Stemming*

Stemming merupakan proses pengolahan teks untuk mengubah setiap kata yang memiliki imbuhan menjadi kata dasar (Kamil et al., 2025). Tahapan ini dilakukan setelah proses *stopword* agar kata yang tersisa diubah ke dalam bentuk dasar. Tujuannya agar meningkatkan pemrosesan teks dalam tahapan algoritma *machine learning* dan pengambilan informasi. Proses ini menggunakan *library python* Sastrawi berbahasa Indonesia khusus *stemmer* digunakan untuk merubah kata yang berimbungan menjadi bentuk kata dasar dalam bahasa Indonesia (Himawan & Eliyani, 2021). Berikut contoh penerapan *stemming* dalam data teks:

Tabel 3.7 Contoh Tahapan Stemming

Stopword Removal	Stemming
[gila, banget, orang, pertamina, rakus, uang, ini]	[gila, banget, orang, pertamina, rakus, uang]
[coba, ferry, tanya, ahok, membiarkan, korupsi, pertamina]	[coba, ferry, tanya, ahok, biar, korupsi, pertamina]
[paling, jahat, korupsi, haji, bang]	[paling, jahat, korupsi, haji, bang]
[hari, terima, kabar, buruk, negara]	[hari, terima, kabar, buruk, negara]
[dengar, jadi, tambah, emosi]	[dengar, jadi, tambah, emosi]
[catet, hakimnya, hujat, terror, selidiki, sutradaranya, cari, temukan]	[catet, hakim, hujat, terror, selidik, sutradara, cari, temu]
[ekonomi, kapitalis, biasanya, disukai, komunis]	[ekonomi, kapitalis, biasa, suka, komunis]
[kualitas, penegak, hukum, buruk, sekali, iya]	[kualitas, tegak, hukum, buruk, sekali, iya]
[keadilan, mati]	[adil, mati]
[indonesia, jadi, hakim, asal, mau, nurut, sama, penguasa]	[indonesia, jadi, hakim, asal, mau, nurut, sama, kuasa]

3.3.4 Pembobotan TF-IDF

Pembobotan dilakukan untuk menghitung kemunculan kata pada setiap dokumen. Tahapan ini bertujuan untuk mengubah teks hasil *processing* menjadi nilai numerik. Terdapat beberapa proses dalam tahapan pembobotan kata menggunakan metode TF-IDF (*Term Frequency-Inverse Document Frequency*), merupakan metode pembobotan kata dihitung berdasarkan kemunculan kata dalam suatu dokumen. Melalui metode ini, setiap kata diberikan nilai bobot sesuai jumlah kemunculan dan keunikannya. Pertama melakukan perhitungan terhadap kemunculan kata pada dokumen, contohnya sebagai berikut:

Tabel 3.8 Contoh Tahapan Term Frequency

Term	**Perhitungan *Term Frequency***									
	D1	**D2**	**D3**	**D4**	**D5**	**D6**	**D7**	**D8**	**D9**	**D10**
adil	0	0	0	0	0	0	0	0	1	0
ahok	0	1	0	0	0	0	0	0	0	0
asal	0	0	0	0	0	0	0	0	0	1
bang	0	0	1	0	0	0	0	0	0	0
banget	1	0	0	0	0	0	0	0	0	0
biar	0	1	0	0	0	0	0	0	0	0
buruk	0	0	0	1	0	0	0	1	0	0
cari	0	0	0	0	0	1	0	0	0	0
catet	0	0	0	0	0	1	0	0	0	0
coba	0	1	0	0	0	0	0	0	0	0
dengar	0	0	0	0	1	0	0	0	0	0
ekonomi	0	0	0	0	0	0	1	0	0	0
emosi	0	0	0	0	1	0	0	0	0	0
ferry	0	1	0	0	0	0	0	0	0	0
gila	1	0	0	0	0	0	0	0	0	0
haji	0	0	1	0	0	0	0	0	0	0
hakim	0	0	0	0	0	1	0	0	0	1
hari	0	0	0	1	0	0	0	0	0	0
hukum	0	0	0	0	0	0	0	1	0	0
hujat	0	0	0	0	0	1	0	0	0	0
indonesia	0	0	0	0	0	0	0	0	0	1
iya	0	0	0	0	0	0	0	1	0	0
jadi	0	0	0	0	1	0	0	0	0	1
jahat	0	0	1	0	0	0	0	0	0	0
kabar	0	0	0	1	0	0	0	0	0	0
kapitalis	0	0	0	0	0	0	1	0	0	0
komunis	0	0	0	0	0	0	1	0	0	0
korupsi	0	1	1	0	0	0	0	0	0	0
kualitas	0	0	0	0	0	0	0	1	0	0

Term	Perhitungan Term Frequency									
	D1	**D2**	**D3**	**D4**	**D5**	**D6**	**D7**	**D8**	**D9**	**D10**
kuasa	0	0	0	0	0	0	0	0	0	1
mati	0	0	0	0	0	0	0	0	1	0
mau	0	0	0	0	0	0	0	0	0	1
negara	0	0	0	1	0	0	0	0	0	0
nurut	0	0	0	0	0	0	0	0	0	1
orang	1	0	0	0	0	0	0	0	0	0
paling	0	0	1	0	0	0	0	0	0	0
pertamina	1	1	0	0	0	0	0	0	0	0
rakus	1	0	0	0	0	0	0	0	0	0
sama	0	0	0	0	0	0	0	0	0	1
sekali	0	0	0	0	0	0	0	1	0	0
selidik	0	0	0	0	0	1	0	0	0	0
suka	0	0	0	0	0	0	1	0	0	0
sutradara	0	0	0	0	0	1	0	0	0	0
tambah	0	0	0	0	1	0	0	0	0	0
tanya	0	1	0	0	0	0	0	0	0	0
tegak	0	0	0	0	0	0	0	1	0	0
terima	0	0	0	1	0	0	0	0	0	0
terror	0	0	0	0	0	1	0	0	0	0
temu	0	0	0	0	0	1	0	0	0	0
uang	1	0	0	0	0	0	0	0	0	0
panjang dokumen	6	7	5	5	4	8	5	6	2	8

Pada tabel 3.8 setiap kata dihitung kemunculannya pada setiap dokumen. Proses ini penting agar dokumen yang memiliki jumlah kata lebih banyak tidak secara otomatis menghasilkan nilai bobot yang lebih tinggi dibandingkan dokumen yang lebih pendek. Berikut contoh penerapan TF normalisasi:

Tabel 3.9 Contoh Tahapan TF Normalisasi

Term	***Term Frequency* Normalisasi**									
	D1	**D2**	**D3**	**D4**	**D5**	**D6**	**D7**	**D8**	**D9**	**D10**
adil	0	0	0	0	0	0	0	0	0,167	0
ahok	0	0,167	0	0	0	0	0	0	0	0
asal	0	0	0	0	0	0	0	0	0	0,167
bang	0	0	0,167	0	0	0	0	0	0	0
banget	0,167	0	0	0	0	0	0	0	0	0
biar	0	0,167	0	0	0	0	0	0	0	0
biasa	0	0	0	0	0	0	0,167	0	0	0
buruk	0	0	0	0,167	0	0	0	0,167	0	0
cari	0	0	0	0	0	0,167	0	0	0	0
catet	0	0	0	0	0	0,167	0	0	0	0
coba	0	0,167	0	0	0	0	0	0	0	0
dengar	0	0	0	0	0,167	0	0	0	0	0
ekonomi	0	0	0	0	0	0	0,167	0	0	0
emosi	0	0	0	0	0,167	0	0	0	0	0
ferry	0	0,167	0	0	0	0	0	0	0	0
gila	0,167	0	0	0	0	0	0	0	0	0

Term	*Term Frequency* **Normalisasi**									
	D1	**D2**	**D3**	**D4**	**D5**	**D6**	**D7**	**D8**	**D9**	**D10**
haji	0	0	0,167	0	0	0	0	0	0	0
hakim	0	0	0	0	0	0,167	0	0	0	0,167
hari	0	0	0	0,167	0	0	0	0	0	0
hukum	0	0	0	0	0	0	0	0,167	0	0
hujat	0	0	0	0	0	0,167	0	0	0	0
indonesia	0	0	0	0	0	0	0	0	0	0,167
iya	0	0	0	0	0	0	0	0,167	0	0
jadi	0	0	0	0	0,167	0	0	0	0	0,167
jahat	0	0	0,167	0	0	0	0	0	0	0
kabar	0	0	0	0,167	0	0	0	0	0	0
kapitalis	0	0	0	0	0	0	0,167	0	0	0
komunis	0	0	0	0	0	0	0,167	0	0	0
korupsi	0	0,167	0,167	0	0	0	0	0	0	0
kualitas	0	0	0	0	0	0	0	0,167	0	0
kuasa	0	0	0	0	0	0	0	0	0	0,167
mati	0	0	0	0	0	0	0	0	0,167	0
mau	0	0	0	0	0	0	0	0	0	0,167

Term	***Term Frequency* Normalisasi**									
	D1	**D2**	**D3**	**D4**	**D5**	**D6**	**D7**	**D8**	**D9**	**D10**
negara	0	0	0	0,167	0	0	0	0	0	0
nurut	0	0	0	0	0	0	0	0	0	0,167
orang	0,167	0	0	0	0	0	0	0	0	0
paling	0	0	0,167	0	0	0	0	0	0	0
pertamina	0,167	0,167	0	0	0	0	0	0	0	0
rakus	0,167	0	0	0	0	0	0	0	0	0
sama	0	0	0	0	0	0	0	0	0	0,167
sekali	0	0	0	0	0	0	0	0,167	0	0
selidik	0	0	0	0	0	0,167	0	0	0	0
suka	0	0	0	0	0	0	0,167	0	0	0
sutradara	0	0	0	0	0	0,167	0	0	0	0
tambah	0	0	0	0	0,167	0	0	0	0	0
tanya	0	0,167	0	0	0	0	0	0	0	0
tegak	0	0	0	0	0	0	0	0,167	0	0
terima	0	0	0	0,167	0	0	0	0	0	0
terror	0	0	0	0	0	0,167	0	0	0	0

Pada tabel 3.9 dilakukan perhitungan normalisasi agar dokumen yang lebih panjang tidak secara otomatis dianggap lebih penting. Contoh perhitungan TF Normalisasi:

$$\boldsymbol{TF(banget) = \frac{count(1)}{6} = 0,167} \tag{3.1}$$

Selanjutnya melakukan penjumlahan *Document Frequency* (DF) dan *Inverse Documen Frequency* (IDF) mengukur keunikan atau kelangkaan suatu *term* pada dokumen. Contoh penerapannya sebagai berikut:

Tabel 3.10 Contoh Tahapan DF dan IDF

Term	DF	IDF
adil	1	1
ahok	1	1
asal	1	1
bang	1	1
banget	1	1
biar	1	1
biasa	1	1
buruk	2	0,699
cari	1	1
catet	1	1
coba	1	1
dengar	1	1
ekonomi	1	1
emosi	1	1
ferry	1	1
gila	1	1
haji	1	1
hakim	2	0,699
hari	1	1
hukum	1	1
hujat	1	1
indonesia	1	1

Term	DF	IDF
iya	1	1
jadi	2	0,699
jahat	1	1
kabar	1	1
kapitalis	1	1
komunis	1	1
korupsi	2	0,699
kualitas	1	1
kuasa	1	1
mati	1	1
mau	1	1
negara	1	1
nurut	1	1
orang	1	1
paling	1	1
pertamina	2	0,699
rakus	1	1
sama	1	1
sekali	1	1
selidik	1	1
suka	1	1
sutradara	1	1
tambah	1	1
tanya	1	1
tegak	1	1
terima	1	1
terror	1	1
temu	1	1
uang	1	1

Pada tabel 3.10 dilakukan perhitungan IDF untuk mengukur seberapa penting atau unik sebuah kata di dalam seluruh dokumen. Contoh perhitungan IDF:

$$\boldsymbol{IDF((pertamina)) = log\frac{(10)}{2} = log(5) = 0,699} \quad (3.2)$$

Pada perhitungan *term* pertamina yang dilakukan pada proses IDF menghasilkan ***log(5)*** yang berarti hasilnya 0,699. Selanjutnya menghitung bobot TF-IDF dengan mengkalikan nilai TF (*Term Frequency*) dengan IDF (*Inverse Document Frequency*). Berikut contoh penerapan TF-IDF:

Tabel 3.11 Contoh Tahapan Pembobotan TF-IDF

Term	***Term Frequency Inverse Document Frequency***									
	D1	**D2**	**D3**	**D4**	**D5**	**D6**	**D7**	**D8**	**D9**	**D10**
adil	0	0	0	0	0	0	0	0	0,167	0
ahok	0	0,167	0	0	0	0	0	0	0	0
asal	0	0	0	0	0	0	0	0	0	0,167
bang	0	0	0,167	0	0	0	0	0	0	0
banget	0,167	0	0	0	0	0	0	0	0	0
biar	0	0,167	0	0	0	0	0	0	0	0
biasa	0	0	0	0	0	0	0,167	0	0	0
buruk	0	0	0		0	0	0	0,117	0	0
cari	0	0	0	0	0	0,167	0	0	0	0
catet	0	0	0	0	0	0,167	0	0	0	0
coba	0	0,167	0	0	0	0	0	0	0	0
dengar	0	0	0	0	0,167	0	0	0	0	0

Term	***Term Frequency Inverse Document Frequency***									
	D1	**D2**	**D3**	**D4**	**D5**	**D6**	**D7**	**D8**	**D9**	**D10**
ekonomi	0	0	0	0	0	0	0,167	0	0	0
emosi	0	0	0	0	0,167	0	0	0	0	0
ferry	0	0,167	0	0	0	0	0	0	0	0
gila	0,167	0	0	0	0	0	0	0	0	0
haji	0	0	0,167	0	0	0	0	0	0	0
hakim	0	0	0	0	0	0,117	0	0	0	0,117
hari	0	0	0	0,167	0	0	0	0	0	0
hukum	0	0	0	0	0	0	0	0,167	0	0
hujat	0	0	0	0	0	0,167	0	0	0	0
indonesia	0	0	0	0	0	0	0	0	0	0,167
iya	0	0	0	0	0	0	0	0,167	0	0
jadi	0	0	0	0	0,117	0	0	0	0	0,117
jahat	0	0	0,167	0	0	0	0	0	0	0
kabar	0	0	0	0,167	0	0	0	0	0	0
kapitalis	0	0	0	0	0	0	0,167	0	0	0
komunis	0	0	0	0	0	0	0,167	0	0	0
korupsi	0	0,117	0,117	0	0	0	0	0	0	0

Term	***Term Frequency Inverse Document Frequency***									
	D1	**D2**	**D3**	**D4**	**D5**	**D6**	**D7**	**D8**	**D9**	**D10**
kualitas	0	0	0	0	0	0	0	0,167	0	0
kuasa	0	0	0	0	0	0	0	0	0	0,167
mati	0	0	0	0	0	0	0	0	0,167	0
negara	0	0	0	0,167	0	0	0	0	0	0
nurut	0	0	0	0	0	0	0	0	0	0,167
orang	0,167	0	0	0	0	0	0	0	0	0
paling	0	0	0,167	0	0	0	0	0	0	0
pertamina	0,117	0,117	0	0	0	0	0	0	0	0
rakus	0,167	0	0	0	0	0	0	0	0	0
sama	0	0	0	0	0	0	0	0	0	0,167
sekali	0	0	0	0	0	0	0	0,167	0	0
selidik	0	0	0	0	0	0,167	0	0	0	0
suka	0	0	0	0	0	0	0,167	0	0	0
sutradara	0	0	0	0	0	0,167	0	0	0	0
tambah	0	0	0	0	0,167	0	0	0	0	0
tanya	0	0,167	0	0	0	0	0	0	0	0
tegak	0	0	0	0	0	0	0	0,167	0	0

Term	*Term Frequency Inverse Document Frequency*									
	D1	**D2**	**D3**	**D4**	**D5**	**D6**	**D7**	**D8**	**D9**	**D10**
terima	0	0	0	0,167	0	0	0	0	0	0

Pada tabel 3.11 dilakukan perhitungan TF-IDF untuk menemukan kata paling relevan mewakili isi dokumen. Contoh perhitungan TF-IDF:

$$\mathbf{W(buruk\ di\ D4)} = \boldsymbol{TF(buruk\ di\ D4)} \times \boldsymbol{IDF(buruk)} \quad (3.3)$$

$$\mathbf{W(buruk\ di\ D4)} = \mathbf{0,167 \times 0,699}$$

$$\mathbf{W(buruk\ di\ D4)} = \mathbf{0,117}$$

3.3.5 Splitting Data

Splitting Data merupakan tahapan pembagian data untuk pelatihan (*training*) dan pengujian (*testing*) pada data teks yang ditentukan oleh rasio dalam proses ini. Rasio yang digunakan sangat bervariatif, umumnya rasio 70:30, 80:20, dan 90:10 (Harianto et al., 2025). Pada penelitian saat ini rasio yang digunakan untuk membagi data latih dan data uji sebesar 70:30 dan 80:20, di mana masing-masing rasio yaitu 70% dan 80% sebagai data latih. Sedangkan, 30% dan 20% sebagai data uji.

3.3.6 Klasifikasi Random Forest

Setelah data teks komentar berhasil diubah menjadi bentuk numerik melalui tahapan pembobotan kata dengan TF-IDF, selanjutnya data teks dilakukan pengklasifikasian menggunakan algoritma klasifikasi *random forest*. Algoritma ini merupakan metode *ensemble learning* yang menggabungkan beberapa pohon keputusan (*decision tree*) untuk menghasilkan prediksi yang akurat (Nugroho, 2025). Terdapat beberapa tahapan untuk membentuk sebuah pohon keputusan dalam algoritma *Random Forest*, seperti melakukan pembagian data latih dan data uji, membuat dataset untuk masing-masing pohon dengan

proses *bagging* (*Bootstrap Aggregating*), penentuan *node* pada pohon keputusan menggunakan *gini impurity*, dan menentukan hasil dengan sistem *majority voting*.

Tabel 3.12 Contoh Split Data Latih dan Data Uji

Rasio	Data Latih	Data Uji
70:30	7	3
80:20	8	2

Pada tabel 3.12 Pada penelitian ini menggunakan sampel sebanyak 10 dokumen yang sudah melalui tahapan *preprocessing* dan pembobotan kata, untuk pembagian data diambil sebagai contoh rasion 80:20 dengan perbandingan 8 sebagai data latih dan 2 sebagai data uji. Pembagian ini dilakukan agar model dapat belajar dari sebagian data, sementar data lainnya digunakan sebagai uji model.. Selain itu pemilihan 80:20 berdasarkan proporsi kemampuan model mempelajari data klasifikasi sentimen. Di mana hasil klasifikasi ini dapat memberikan gambaran akurat mengenai kemampuan algoritma dalam mengklasifikasikan data sentimen.

Tabel 3.13 Contoh Dataset Utama Ratio 80:20

Isi Dokumen	Label
[catet, hakim, hujat, terror, selidik, sutradara, cari, temu]	Negatif
[gila, banget, orang, pertamina, rakus, uang]	Positif
[kualitas, tegak, hukum, buruk, sekali, iya]	Positif
[paling, jahat, korupsi, haji, bang]	Netral
[indonesia, jadi, hakim, asal, nurut, sama, kuasa]	Positif
[dengar, jadi, tambah, emosi]	Positif
[hari, terima, kabar, buruk, negara]	Positif
[ekonomi, kapitalis, biasa, suka, komunis]	Negatif
[adil, mati]	Positif
[coba, ferry, tanya, ahok, biar, korupsi, pertamina]	Netral

Pada tabel 3.13 merupakan gambaran dataset utama dengan rasio 80:20 yang nantinya akan digunakan sebagai perbandingan hasil

prediksi dari model. Dataset ini terdiri dari 10 sampel yang sudah melepati tahapan *preporcessing*, sebagai upaya dalam pembersihan teks dari *noise* dan membuat data lebih konsisten. Selanjutnya, membuat dataset baru untuk masing-masing pohon keputusan dengan metode *bagging*. Contoh pembuatan dataset baru untuk pohon keputusan 1:

Tabel 3.14 Dataset Hasil Bagging Pohon Keputusan 1

Isi Dokumen	Label
[gila, banget, orang, pertamina, rakus, uang]	Positif
[ekonomi, kapitalis, biasa, suka, komunis]	Negatif
[indonesia, jadi, hakim, asal, nurut, sama, kuasa]	Positif
[gila, banget, orang, pertamina, rakus, uang]	Positif
[hari, terima, kabar, buruk, negara]	Positif
[catet, hakim, hujat, terror, selidik, sutradara, cari, temu]	Negatif
[paling, jahat, korupsi, haji, bang]	Netral
[ekonomi, kapitalis, biasa, suka, komunis]	Negatif

Pada tabel 3.14 merupakan dataset baru yang dibuat dengan proses *bagging* (*bootstrap aggregating*). Selanjutnya menentukan simpul atau *node* dari pohon keputusan 1 menggunakan perhitungan *gini impurity* untuk mengukur kemurnian kata dalam memisahkan data berdasarkan kelas-kelasnya.

$$\boldsymbol{Gini = 1 - \sum_{i=1}^{c} (pi)^2} \tag{3.4}$$

$$Giniparent(1) = 1 - \left[\left(\frac{4}{8}\right)^2 + \left(\frac{1}{8}\right)^2 + \left(\frac{3}{8}\right)^2\right]$$

$$Giniparent(1) = 1 - 0{,}40625 = 0{,}59375$$

Pada rumus 3.4 merupakan rumus *Gini* untuk menghitung tingkat keragaman (*impurity*) dari seuluruh dataset pohon keputusan 1. Di mana *c* menunjukan jumlah kelas (positif, negatif, netral), pi menunjukan presentasi data kelas ke-i. Jadi terdapat 4 kelas positif, 1 kelas negatif, dan 3 kelas netral yang dibagi masing-masing dengan total dokumen yaitu 8. Kemudian hasil akhirnya, 0,59375 merupakan

tingkat keberagaman awal. Selanjutnya *split data* menjadi dua grup (kiri dan kanan).

Tabel 3.15 Contoh Split Data

Kategori	Jumlah	Isi
Data kiri (*True*)	6	4 Positif, 1 Negatif, 1 Netral
Data kanan (*False*)	2	2 Negatif

Pada tabel 3.15 merupakan *split data* kiri dan kanan yang mewakili sebelah kiri yang mengandung tidak mengandung kata 'biasa' dan sebelah kanan mengandung kata 'biasa'.

$$Ginikiri(1) = 1 - \left[\left(\frac{4}{6}\right)^2 + \left(\frac{1}{6}\right)^2 + \left(\frac{1}{6}\right)^2\right] = 1 - 0{,}5 = 0{,}5 \quad (3.5)$$

$$Ginikanan(1) = 1 - \left[\left(\frac{2}{2}\right)^2\right] = 0$$

Pada rumus 3.5 dilakukan perhitungan *gini* terhadap masing-masing kategori. Tujuannya untuk mengukur tingkat keragaman setelah di split pada kategori kiri dan kanan. Selanjutnya menghitung rata-rata gini kiri dan kanan:

$$\boldsymbol{Ginisplit} = \left(\frac{\boldsymbol{nkiri}}{\boldsymbol{ntotal}} \times \boldsymbol{Ginikiri}\right) + \left(\frac{\boldsymbol{nkanan}}{\boldsymbol{ntotal}} \times \boldsymbol{Ginikanan}\right) \quad (3.6)$$

$$Ginisplit(1) = \left(\frac{6}{8} \times 0{,}5\right) + \left(\frac{2}{8} \times 0\right) = 0{,}75 \times 0{,}5 + 0 = 0{,}375$$

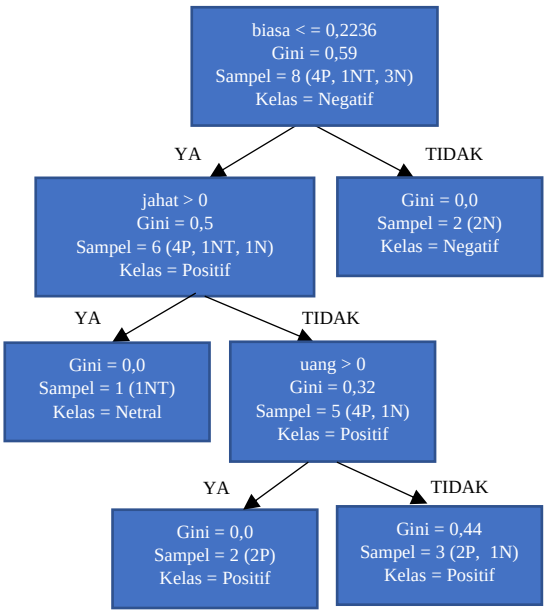
Pada rumus 3.6 dilakukan perhitungan *gini split* dari kedua kategori, di dapatkan hasilnya 0,375 sebagai rata-rata dari perhitungan keduanya. Selanjutnya menghitung gini gain sebagai berikut:

$$\boldsymbol{Ginigain} = \boldsymbol{Giniparent} - \boldsymbol{Ginisplit} \quad (3.7)$$

$$Ginigain(1) = 0{,}59375 - 0{,}375 = 0{,}21875$$

Pada rumus 3.7 dilakukan perhitungan *gini gain* untuk mengukur seberapa efektif *split* yang dilakukan dengan membandingkan tingkat keragaman awal dan tingkat keragaman rata-rata setelah *split*. Di

dapatkan hasilnya 0,21875 sebagai tingkat penurunan keragaman. Tahap selanjutnya setelah menemukan *node*, membuat alur pohon keputusan 1 seperti berikut:

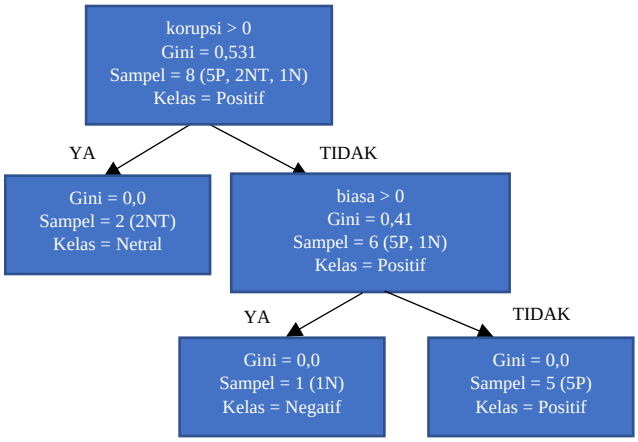


Gambar 3.3 Pohon Keputusan 1

(*Sumber: Hasil Olah Data*)

Pada gambar 3.3 memvisualisasikan struktur dari pohon keputusan pertama yang merupakan komponen integral dalam pembentukan model *Random Forest*. Dalam skema ini, fitur kata "biasa" ditetapkan sebagai simpul akar, karena kata tersebut memiliki nilai *Gini Gain* paling signifikan sebesar 0,59 dibandingkan fitur lainnya. Pada titik awal, terdapat delapan sampel data yang diproses dengan distribusi label yang terdiri dari empat komentar positif, satu komentar netral, dan tiga komentar negatif. Struktur pohon dibagi menjadi dua jalur utama, apabila terdeteksi input fitur "biasa", algoritma otomatis mengarahkannya ke arah kanan dengan klasifikasi negatif. Sebaliknya, jika tidak mengandung fitur tersebut, alur akan ke arah kiri untuk dilakukan pemisahan lanjutan dengan fitur "jahat" dan "uang". Dengan serangkaian evaluasi pada simpul-simpul, pohon

keputusan akhirnya mampu menghasilkan keputusan akhir secara konsisten ke dalam klasifikasi positif.

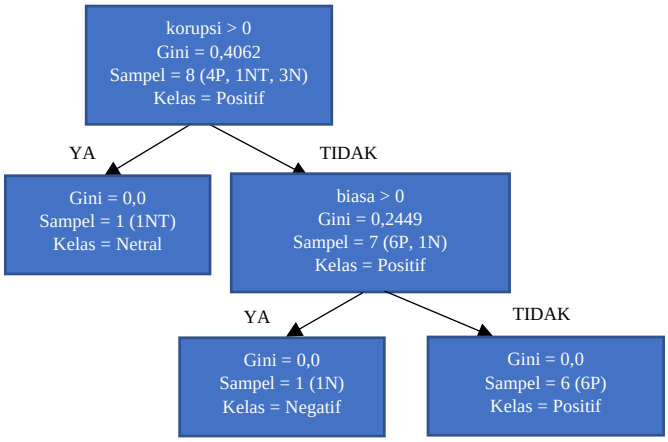


Gambar 3.4 Pohon Keputusan 2

(*Sumber: Hasil Olah Data*)

Pada gambar 3.4 merupakan struktur pohon keputusan 2 yang dibentuk dari simpul kata dalam algoritma *Random Forest*. Pada penelitian ini, fitur kata "korupsi"dipilih karena memiliki nilai *gini gain* tertinggi dengan nilai 0,531. Simpul akar ini memproses total delapan sampel data yang distribusinya berupa lima label positif, dua label netral, dan satu label negatif. Mekanisme pengambilan keputusan pada tahap pertama menunjukkan, apabila data input mengandung fitur kata "korupsi" maka melewati jalur YA, alur prediksi langsung menuju simpul daun dengan nilai *Gini* 0,0 yang mengindikasikan kemurnian label netral sebanyak dua sampel. Sebaliknya, jika fitur tidak ditemukan dalam data input, maka proses melewati jalur TIDAK proses akan bergeser pada simpul internal berikutnya yang mengandung fitur kata "biasa" dengan nilai *Gini* 0,41, maka akan terbentuk simpul negatif untuk sampel. Namun, apabila tidak terdeteksi ke dua simpul sebelumnya, secara otomatis model akan memasukkan data ke dalam klasifikasi label positif dengan nilai *Gini* 0,0. Dengan melalui struktur pohon keputusan ini mampu menangkap variasi pola sentimen,

sehingga secara keseluruhan mendukung peningkatan hasil akhir klasifikasi model.



Gambar 3.5 Pohon Keputusan 3

(*Sumber: Hasil Olah Data*)

Pada gambar 3.5 merupakan struktur pohon keputusan 3 yang dibentuk dari fitur utama kata “korupsi” karena memiliki nilai *gini gain* tertinggi dengan nilai 0,41 dengan delapan sampel terdiri dari empat positif, satu netral, dan tiga negatif yang mayoritas kelasnya positif. Cabang yang mengandung kata “korupsi” menghasilkan node murni kelas netral, sedangkan cabang tanpa kata “korupsi” dilakukan pemisahan lanjutan dengan fitur “biasa” dengan *gini* 0,24. Pada percabangan selanjutnya, jika terdapat kata “biasa” maka terbentuk *node* negatif dengan satu sampel. Sedangkan tidak terdapat kata “biasa” terbentuk *node* positif dengan enam sampel.

Pada tahap selanjutnya dilakukan pengujian hasil prediksi masing-masing pohon keputusan pada data uji. Tujuannya untuk melihat kemampuan model dalam memprediksi kelas sentimen dari dokumen baru berdasarkan pelatihan pada data latih. Berikut hasil prediksi model:

Tabel 3.16 Contoh Pengujian Data

Isi Dokumen	Label *Actual*	Label *Predicted*
[adil, mati]	Positif	Positif
[coba, ferry, tanya, ahok, biar, korupsi, pertamina]	Netral	Positif

Pada tabel 3.16 menunjukkan hasil penerapan data uji terhadap ketiga model pohon keputusan. Di mana setiap pohon keputusan diuji menggunakan dokumen yang berisi kata-kata sesuai fitur utama. Pada dokumen pertama (“adil, mati”) model memprediksi kelas positif karena tidak ditemukan kata dengan kecenderungan negatif. Selanjutnya pada dokumen kedua (“tanya, ahok, biar, korupsi, pertamina”) kemunculan kata “korupsi” membuat model memprediksi positif karena mengandung salah satu kata pada fitur utama pohon keputusan.

3.3.7 Confusion Matrix

Setelah data dokumen berhasil di prediksi oleh model klasifikasi *random forest,* selanjutnya model akan di evaluasi menggunakan *confusion matrix*. *Confusion matrix* merupakan matriks evaluasi berbentuk sebuah tabel yang berisi kolom ‘*actual*’ dan ‘*prediction*’ fungsinya untuk membandingkan hasil klasifikasi oleh model dengan kategori klasifikasi sebenarnya (Amin, 2023). Berikut penerapan *confusion matrix* pada hasil prediksi model klasifikasi *random forest*:

Tabel 3.17 Contoh Tabel Confusion Matrix

Aktual				
Positif	Netral	Negatif		
1	0	0	Positif	Prediksi
1	0	0	Netral	
0	0	0	Negatif	

Pada tabel 3.17 terdapat sebuah tabel matriks perbandingan antara nilai aktual dan nilai prediksi. Tujuannya untuk mengukur kinerja

model dalam memprediksi data uji dengan benar, terlihat bahwa dari hasil prediksi pada data uji didapatkan hasil prediksi yang benar pada kelas positif dan kurang tepat pada kelas netral. Kemudian dituliskan setiap nilai dari hasil prediksi ke dalam tabel, nantinya akan dilakukan pengukuran berdasarkan matriks *accuracy*, *precision*, *recall*, dan *f1-score* seperti berikut:

Perhitungan *Accuracy*

$$\boldsymbol{Accuracy} = \frac{(\boldsymbol{Jumlah\ Prediksi\ Benar})}{(\boldsymbol{Total\ Data})} \tag{3.8}$$

$$\boldsymbol{Accuracy} = \frac{\mathbf{1+0+0}}{\mathbf{1+1++0}} = \frac{\mathbf{1}}{\mathbf{2}} = \mathbf{0,5}$$

Hasil Akurasi = 50%

Pada rumus 3.8 merupakan rumus menghitung akurasi sebuah model dalam memprediksi kelas sesuai dengan label aktualnya. Diketahui jumlah prediksi benar hanya satu pada kelas positif dari total dua data uji. Hasil yang diperoleh model sebesar 0,5 atau 50%, artinya model hanya mampu memprediksi dengan benar hanya setengahnya dari total data uji.

Perhitungan *Precision*

$$\boldsymbol{Precision} = \frac{(\boldsymbol{TP})}{(\boldsymbol{TP+FP})} \tag{3.9}$$

$$\boldsymbol{Precision}(\boldsymbol{Positif}) = \frac{\mathbf{1}}{\mathbf{1+0}} = \mathbf{1}$$

$$\boldsymbol{Precision}(\boldsymbol{Netral}) = \frac{\mathbf{0}}{\mathbf{0+1}} = \mathbf{0}$$

$$\boldsymbol{Precision}(\boldsymbol{Negatif}) = \frac{\mathbf{0}}{\mathbf{0+0}} = \mathbf{0}$$

Pada rumus 3.9 merupakan rumus menghitung presisi untuk menghitung ketepatan model dalam mengklasifikasikan data sesuai dengan kelasnya. Berdasarkan perhitungan, nilai presisi kelas positif hanya satu sesuai dengan label aktual pada data uji. Sementara itu, nilai

presisi kelas netral dan negatif tidak mendapatkan nilai karena model gagal mengklasifikasikan data ke dalam dua kelas tersebut.

Perhitungan *Recall*

$$\boldsymbol{Recall} = \frac{(\boldsymbol{TP})}{(\boldsymbol{TP} + \boldsymbol{FN})} \tag{3.10}$$

$$\boldsymbol{Recall(Positif)} = \frac{\mathbf{1}}{\mathbf{1}+\mathbf{1}} = \mathbf{0,5}$$

$$\boldsymbol{Recall(Netral)} = \frac{\mathbf{0}}{\mathbf{0}+\mathbf{1}} = \mathbf{0}$$

$$\boldsymbol{Recall(Negatif)} = \frac{\mathbf{0}}{\mathbf{0}+\mathbf{0}} = \mathbf{0}$$

Pada rumus 3.10 merupakan rumus menghitung recall untuk menunjukkan kemampuan model dalam mengenali atau menemukan kembali seluruh data yang sesuai dengan kelas pada label aktual. Recall dihitung dengan membandingkan hasil prediksi benar terhadap total label aktual. Berdasarkan perhitungan, nilai recall kelas positif sebesar 0,5, karena model hanya berhasil menemukan kembali setengah data dari total label aktual kelas positif. Sementara, pada kelas netral dan negatif mendapatkan nilai 0, karena model gagal menemukan kembali seluruh data yang sesuai dengan kelas pada label aktualnya.

Perhitungan *F1-Score*

$$\boldsymbol{f1-score} = \mathbf{2} \times \frac{\boldsymbol{Precision} \times \boldsymbol{Recall}}{\boldsymbol{Precision} + \boldsymbol{Recall}} \tag{3.11}$$

$$\boldsymbol{f1-score(Positif)} = \mathbf{2} \times \frac{\mathbf{1} \times \mathbf{0,5}}{\mathbf{1}+\mathbf{0,5}} = \mathbf{2} \times \frac{\mathbf{0,5}}{\mathbf{1,5}} = \mathbf{2} \times \frac{\mathbf{1}}{\mathbf{1,5}}$$

$$= \mathbf{0,667}$$

$$\boldsymbol{f1-score(Netral)} = \mathbf{2} \times \frac{\mathbf{0} \times \mathbf{0}}{\mathbf{0}+\mathbf{0}} = \mathbf{0}$$

$$\boldsymbol{f1-score(Negatif)} = \mathbf{2} \times \frac{\mathbf{0} \times \mathbf{0}}{\mathbf{0}+\mathbf{0}} = \mathbf{0}$$

Pada rumus 3.11 merupakan rumus menghitung *f1-score* untuk mengukur keseimbangan model antara nilai *precision* dan *recall* pada

setiap kelas. Tujuannya memberikan *insight* terhadap kemampuan model dalam memprediksi akurat dan konsisten dalam mengenali data. Berdasarkan perhitungan, nilai *f1-score* kelas positif sebesar 0,667, menunjukkan model memiliki kinerja yang baik dalam memprediksi kelas positif yang benar. Sedangkan pada kelas netral dan negatif tidak mendapatkan nilai, karena model gagal dalam memprediksi kedua kelas ini.

3.4 Tempat dan Waktu Penelitian

Penelitian ini berfokus terhadap topik korupsi yang dibahas di kanal YouTube Malaka Project. Penelitian ini dilakukan pada tanggal 27 Mei 2025 hingga April 2026. Tahapan penelitian diawali dengan penentuan topik penelitian di bulan Mei 2025, pengumpulan data atau *scraping* dilakukan di bulan Oktober 2025, pelabelan data di bulan November 2025 hingga februari 2026, pemilihan algoritma klasifikasi *Random Forest* di bulan Maret 2026, dan evaluasi hasil klasifikasi model di bulan April 2026.

3.5 Subjek dan Objek Penelitian

Penentuan subjek dan objek dilakukan untuk memastikan data yang dikumpulkan relevan dengan tujuan penelitian. Terdapat subjek dan objek penelitian ini, antara lain:

3.5.1 Subjek Penelitian

Subjek dalam penelitian ini adalah pengguna *platform* YouTube yang berinteraksi di kolom komentar pada ke dua video bertopik korupsi yang di dapat dari kanal edukasi Malaka Project.

3.5.2 Objek Penelitian

Objek penelitian ini berfokus pada dua video dari kanal edukasi Malaka Project yang membahas topik korupsi:

1. Video pertama berjudul “Korupsi Terjahat dan Terstruktur Pertamina” membahas dugaan kasus mega korupsi di Pertamina

yang dinilai sangat terencana karena terjadi dari hulu ke hilir selama kurang lebih 5 tahun. Saat ini, sudah ditetapkan sembilan orang tersangka yang termasuk jajaran petinggi dari Pertamina dan anak perusahaan ini diperkirakan mencapai Rp.192 triliun per tahun, meskipun angka pastinya masih dalam perhitungan pihak berwenang (Irwandi, 2025).

2. Video kedua berjudul "Vonis Tom Lembong Adalah Masalah Publik" membahas kejanggalan vonis pidana penjara 4,5 tahun dan denda hampir Rp.800 juta yang dijatuhkan kepada Tom Lembong terkait kebijakan impor Gula Kristal Mentah (GKM) pada tahun 2015-2016. Pasalnya dugaan yang dijatuhkan sangat tidak masuk akal, karena kebijakan impor saat itu dilakukan berdasarkan kondisi krisis suplai menjelang hari Lebaran, di mana stok gula yang tersedia hanya cukup untuk 1,5 bulan dan produksi dalam negeri tidak mampu memenuhi kebutuhan nasional yang mencapai 5,7 juta ton. Oleh karena itu langkah impor gula, merupakan tindakan yang sangat logis (Irlanie, 2025).

3.6 Sumber Data

Penelitian ini menggunakan data primer berupa komentar penonton dari kedua video yang membahas topik korupsi di kanal YouTube Malaka Project. Sedangkan data sekunder diperoleh dari literatur ilmiah, seperti jurnal, buku, dan *literatur review*.

3.7 Teknik Pengumpulan Data

Penelitian ini mengambil data dari kedua video yang membahas topik korupsi di kanal Malaka Project dengan teknik *Scraping* menggunakan *API* YouTube yang diperoleh dari *Google App Script*.

BAB IV
HASIL DAN PEMBAHASAN

4.1 Hasil Penelitian

Pada bab ini menjelaskan hasil dari implementasi metode algoritma klasifikasi Random Forest yang mengklasifikasikan data komentar video yang membahas topik korupsi pada *kanal* Malaka Project. Terdapat beberapa tahapan, yaitu tahap *Scraping data*, pelabelan data, *preprocessing text*, *split data* dengan rasio 80:20 dan 70:30, *train model* algoritma Random Forest, dan evaluasi model yang diukur berdasarkan nilai *accuracy*, *precision*, *recall*, dan *f1-score*.

4.1.1 Pengambilan Data

Dalam penelitian ini, proses pengambilan data dilakukan menggunakan *Google App Script* dengan memanfaatkan *YouTube Data API V3*. Pengambilan data dilakukan dengan menyusun kode program pada editor *Apps Script* untuk mengekstraksi komentar secara otomatis melalui tautan video YouTube yang menjadi objek penelitian. Pengambilan data ini dilakukan pada tanggal 20 Oktober 2025, sebanyak 4.544 komentar berhasil dikumpulkan dari dua video yang membahas topi korupsi pada kanal YouTube Malaka Project.

4.1.2 Pelabelan Data

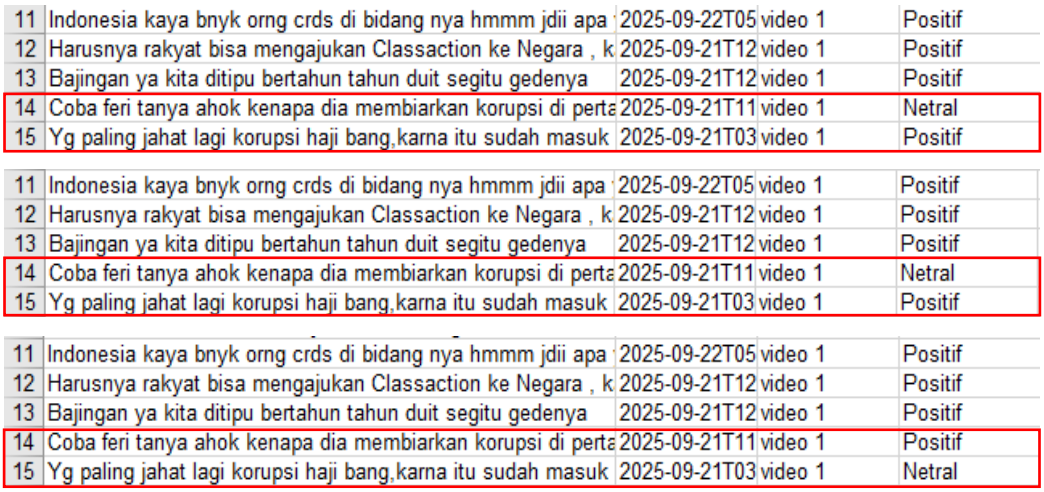
Pada proses pelabelan data menjadi tahapan validasi awal untuk menentukan label yang akan digunakan pada dataset komentar sebagai pedoman dari model Random Forest dalam melakukan klasifikasi. Total keseluruhan data sebanyak 4.544 komentar, yang melibatkan tiga orang ahli bahasa Indonesia. Ketiga ahli tersebut merupakan guru mata pelajaran Bahasa Indonesia yang saat ini aktif mengajar di Yayasan Pondok Pesantren Nurul Islam Mojokerto. Mereka dipilih bukan hanya karena memiliki latar pendidikan formal sebagai Sarjana Pendidikan Bahasa dan

Sastra Indonesia, tetapi juga pengalaman mengajar selama 5 tahun serta pemahaman mendalam terhadap gaya bahasa sehari-hari yang digunakan di media sosial. Selain itu, latar belakang mengajar di pesantren juga membantu dalam memahami etika komunikasi islam dalam mengintepretasikan komentar-komentar yang membahas isu korupsi dari sudut pandang Islam. Pelabelan ini dilakukan pada tanggal 18 Februari 2026.

Proses pelabelan yang dilakukan mengikuti panduan kriteria pelabelan yang sudah dibuat untuk menjamin konsistensi dan objektivitas pengklasifikasian. Setiap komentar pada dataset diklasifikasikan menjadi tiga kelas, yaitu positif, netral, dan negatif. Adapun kriteria masing-masing label seperti berikut. Pada label positif diberikan pada komentar apabila mendukung narasi antikorupsi atau mendukung isi video, menyuarakan penindakan hukum yang tegas terhadap pelaku, mengecam koruptor dan jejaringnya, serta komentar yang menyampaikan terima kasih kepada pembuat konten. Selanjutnya label netral diberikan untuk komentar yang tidak merepresentasikan keberpihakan terhadap narasi antikorupsi, selain itu biasanya berupa candaan, pertanyaan, serta pujian umum tanpa sikap pro maupun kontra. Sementara itu, label negatif diberikan pada komentar yang meremehkan tindakan korupsi, apalagi sampai membela pelaku dan jejaringnya, serta melakukan serangan verbal terhadap pembuat konten.

Tahapan awal dari proses pelabelan, setiap ahli membaca serta menganalisis seluruh komentar secara independent tanpa adanya komunikasi antar ahli. Tujuannya untuk memperoleh intepretasi murni dari masing-masing ahli sehingga meminimalisasikan potensi bias. Tahap kedua, memberikan label (positif, netral, dan negatif) dilakukan dengan mempertimbangkan konteks keseluruhan, seperti kalimat, nada emosional, sarkasme halus, serta budaya yang melekat dalam bahasa komentar masyarakat Indonesia. Seperti penggunaan kata singkatan “KKN” yang diartikan korupsi kolusi dan nepotisme atau ungkapan lainnya yang sering muncul di kolom komentar YouTube pada kedua video ini. Tahap terakhir,

seluruh hasil pelabelan yang sudah diverifikasi oleh masing-masing ahli di dokumentasikan ke dalam file excel dengan format CSV, yang memuat kolom user, teks, date, nomor_video, dan label.



11	Indonesia kaya bnyk orng crds di bidang nya hmmm jdii apa	2025-09-22T05	video 1	Positif
12	Harusnya rakyat bisa mengajukan Classaction ke Negara , k	2025-09-21T12	video 1	Positif
13	Bajingan ya kita ditipu bertahun tahun duit segitu gedenya	2025-09-21T12	video 1	Positif
14	Coba feri tanya ahok kenapa dia membiarkan korupsi di perta	2025-09-21T11	video 1	Netral
15	Yg paling jahat lagi korupsi haji bang,karna itu sudah masuk	2025-09-21T03	video 1	Positif

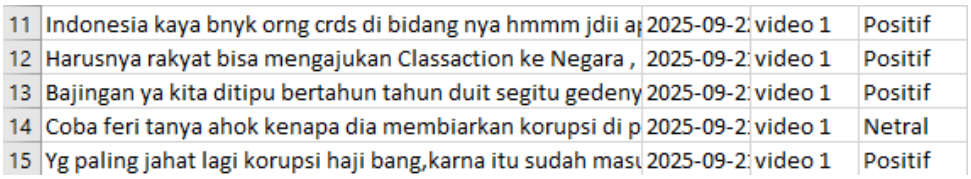
11	Indonesia kaya bnyk orng crds di bidang nya hmmm jdii apa	2025-09-22T05	video 1	Positif
12	Harusnya rakyat bisa mengajukan Classaction ke Negara , k	2025-09-21T12	video 1	Positif
13	Bajingan ya kita ditipu bertahun tahun duit segitu gedenya	2025-09-21T12	video 1	Positif
14	Coba feri tanya ahok kenapa dia membiarkan korupsi di perta	2025-09-21T11	video 1	Netral
15	Yg paling jahat lagi korupsi haji bang,karna itu sudah masuk	2025-09-21T03	video 1	Positif

11	Indonesia kaya bnyk orng crds di bidang nya hmmm jdii apa	2025-09-22T05	video 1	Positif
12	Harusnya rakyat bisa mengajukan Classaction ke Negara , k	2025-09-21T12	video 1	Positif
13	Bajingan ya kita ditipu bertahun tahun duit segitu gedenya	2025-09-21T12	video 1	Positif
14	Coba feri tanya ahok kenapa dia membiarkan korupsi di perta	2025-09-21T11	video 1	Positif
15	Yg paling jahat lagi korupsi haji bang,karna itu sudah masuk	2025-09-21T03	video 1	Netral

Gambar 4.1 Perbandingan Hasil Pelabelan Tiga Ahli Bahasa

(*Sumber: Hasil Olah Data*)

Gambar 4.1 menyajikan contoh perbandingan hasil pelabelan yang dilakukan secara masing-masing oleh ketiga ahli bahasa pada contoh lima komentar yang sama. Kotak merah menyoroti dua komentar yang mengalami perbedaan label. Pada komentar nomor 14, dua ahli memberikan label netral dan satu ahli memberikan positif, sehingga label final ditetapkan sebagai netral. Sementara itu, pada komentar nomor 15, dua ahli memberikan label positif dan satu ahli memberikan netral, sehingga label final ditetapkan sebagai positif. Proses ini menggambarkan penerapan mekanisme pengambilan suara terbanyak untuk menentukan label fix ketika terjadi perbedaan hasil label antar ahli.



11	Indonesia kaya bnyk orng crds di bidang nya hmmm jdii ap	2025-09-2	video 1	Positif
12	Harusnya rakyat bisa mengajukan Classaction ke Negara ,	2025-09-2	video 1	Positif
13	Bajingan ya kita ditipu bertahun tahun duit segitu gedeny	2025-09-2	video 1	Positif
14	Coba feri tanya ahok kenapa dia membiarkan korupsi di p	2025-09-2	video 1	Netral
15	Yg paling jahat lagi korupsi haji bang,karna itu sudah mas	2025-09-2	video 1	Positif

Gambar 4.2 Hasil Pelabelan dari Pemungutan Suara Terbanyak

(*Sumber: Hasil Olah Data*)

Gambar 4.2 menyajikan hasil pelabelan akhir setelah dilakukan pemungutan suara terbanyak oleh ketiga ahli bahasa. Tabel ini menampilkan lima komentar contoh beserta label sentimen akhir yang telah ditetapkan sebagai label fix. Dari lima contoh komentar diatas, empat baris diberikan label positif dan satu baris diberikan label netral. Kotak merah menyoroti dua komentar yang sebelumnya mengalami perbedaan pendapat antar ahli, pada baris 14 dan 15 dengan menerapkan pemungutan suara terbanyak di dapatkan label fix masing-masing baris. Dengan menerapkan mekanisme pemungutan suara terbanyak, label yang diberikan bisa lebih objektif dan dapat dipertanggungjawabkan.

4.1.3 Hasil Preprocessing Text

Pada tahapan selanjutnya, data yang sudah diberi label akan melalui beberapa tahapan persiapan sebelum diolah dengan mesin. Data tersebut perlu dipersiapkan terlebih dahulu melalui *preprocessing text* agar bentuk teks menjadi lebih rapi, konsisten, dan mudah diubah menjadi fitur numerik. *Preprocessing* dilakukan secara sistematis melalui beberapa langkah, seperti *cleaning*, *case folding*, *normalization*, *tokenizing*, *stopwords removal*, dan *stemming*. Selain itu, *preprocessing* juga berfungsi mempertahankan kata-kata yang relevan dan memiliki makna, sehingga teks yang dihasilkan dapat membantu proses ekstraksi fitur dan meningkatkan hasil klasifikasi secara keseluruhan.

Tabel 4.1 Hasil Raw Data

No	*Raw Data*
1	Layak hukuman mati
2	Baru ini korupsi yg dampaknya aku ngerasain bang 😅
3	negeri oplosan
4	Kok ganti kanal
5	Hukum mati bagi maling pertamina..
6	Mereka akan lolos. Kenapa? Uangnya udah banyak

No	*Raw Data*
7	Korupsi lah selagi ada kesempatan.Kalau ketahuan hukumannya ringan kok. Hidup KKN . Taik kucing hukum dinegara konoha
8	Kalau mobil pribadi isi bbm subsidi apakah merugikan negara?
9	Udah males bang liat konten lu, lebih cari aman karena udah jd jongos si botak gak sevokal dulu
10	keren banget baru nemu ini chanel!

Pada tabel 4.1 disajikan contoh data mentah yang diambil dari dataset utama. Data tersebut masih menunjukkan variasi dan inkonsistensi pada teks, seperti penggunaan emoji, tanda baca berulang, spasi berlebih, serta bentuk kata tidak baku. Kondisi tersebut wajar terjadi pada komentar di media sosial, tetapi hal ini dapat mempengaruhi kualitas analisis mesin karena teks yang tidak terstruktur. Jika data mentah langsung diproses ke tahap pembobotan TF-IDF tanpa melalui *preprocessing text*, kemungkinan besar simbol-simbol dan inkonsistensi kata ikut terbaca sebagai fitur, sehingga fitur akan lebih banyak tetapi tanpa berisi informasi penting dari sentimen sehingga menyebabkan model berpotensi bias terhadap pola.

Tabel 4.2 Hasil Tahapan Cleaning

No	*Cleaning*
1	Layak hukuman mati
2	Baru ini korupsi yg dampaknya aku ngerasain bang
3	negeri oplosan
4	Kok ganti kanal
5	Hukum mati bagi maling pertamina
6	Mereka akan lolos Kenapa Uangnya udah banyak
7	Korupsi lah selagi ada kesempatan.Kalau ketahuan hukumannya ringan kok. Hidup KKN Taik kucing hukum dinegara Konoha
8	Kalau mobil pribadi isi bbm subsidi apakah merugikan negara
9	Udah males bang liat konten lu, lebih cari aman karena udah jd jongos si botak gak sevokal dulu
10	keren banget baru nemu ini chanel

Pada tabel 4.2 terlihat bahwa tahap *cleaning* berperan sebagai kontrol kualitas data, tujuannya untuk memastikan teks komentar berisi informasi

yang relevan dengan isi opini, tanpa menyertakan elemen-elemen yang tidak bermakna seperti URL, emoji, mention, tanda baca, spasi berlebih dan lainnya. Tahap ini juga membantu efisiensi fitur yang digunakan pada pelatihan model. *Noise* yang dibiarkan dapat membuat fitur lebih banyak, yang nantinya berpengaruh pada bobot TF-IDF tersebar pada token kata yang tidak bermakna. Hal ini mengakibatkan model membutuhkan daya komputasi lebih besar, selain itu dapat mempengaruhi model dalam mempelajari pola sentimen. Dengan menerapkan tahapan ini pada dataset, komentar menjadi lebih konsisten, lebih mudah diproses pada tahap tokenisasi dan pembobotan kata, serta meningkatkan performa model dalam menangkap kata kunci yang relevan terhadap sentimen.

Tabel 4.3 Hasil Tahapan Case Folding

No	*Case Folding*
1	layak hukuman mati
2	baru ini korupsi yg dampaknya aku ngerasain bang
3	negeri oplosan
4	kok ganti kanal
5	hukum mati bagi maling pertamina
6	mereka akan lolos kenapa angnya udah banyak
7	korupsi lah selagi ada kesempatankalau ketahuan hukumannya ringan kok hidup kkn taik kucing hukum dinegara konoha
8	kalau mobil pribadi isi bbm subsidi apakah merugikan negara
9	udah males bang liat konten lu lebih cari aman karena udah jd jongos si botak gak sevokal dulu
10	keren banget baru nemu ini chanel

Pada tabel 4.3 terlihat tahap *case folding* mengubah seluruh teks menjadi huruf kecil secara konsisten. Tujuannya adalah menyetarakan bentuk kata yang sebenarnya identik, contohnya kata “korupsi”, “Korupsi”, dan “KORUPSI”, agar mesin membaca semua katanya dengan makna sama. Tahap ini sangat berguna sebelum teks dipisah menjadi token-token kata, karena perbedaan huruf kapital dapat membuat kata yang dianggap sama menjadi fitur kata yang berbeda. Dengan menerapkan

tahapan ini, membantu mengurangi variasi dan menekan jumlah fitur kata yang muncul. Hal ini, dapat memperkuat konsistensi data dan memperjelas representasi data.

Tabel 4.4 Hasil Tahapan Normalization

No	*Normalization*
1	layak hukuman mati
2	baru ini korupsi yang dampaknya aku ngerasain bang
3	negeri oplosan
4	kok ganti kanal
5	hukum mati bagi maling pertamina
6	mereka akan lolos kenapa uangnya udah banyak
7	korupsi lah selagi ada kesempatankalau ketahuan hukumannya ringan kok hidup korupsi kolusi nepotisme taik kucing hukum dinegara konoha
8	kalau mobil pribadi isi bbm subsidi apakah merugikan negara
9	udah males bang liat konten lu lebih cari aman karena udah jadi jongos si botak gak sevokal dulu
10	keren banget baru nemu ini chanel

Pada tabel 4.4 disajikan tahap *normalization* yang dilakukan setelah teks dibuat seragam melalui *case folding*. Tahap ini bertujuan menyelaraskan bentuk kata agar sesuai dengan penulisan dan ejaan umum bahasa Indonesia yang lebih baku, serta mengurangi variasi kata yang muncul akibat bahasa gaul, singkatan, atau ejaan tidak konsisten pada komentar di media sosial. Tahapan ini dilakukan dengan mencari kata ganti yang tepat menggunakan pedoman KBBI (Kamus Besar Bahasa Indonesia) yang di inputkan di dalam *script code* pada saat menjalankan program ini. Tahapan ini sangat penting dilakukan karena model belajar memahami pola berdasarkan token-token kata. Jika terdapat banyak variasi kata, dapat mempengaruhi bobot ekstraksi fitur menjadi lemah. Dengan menerapkan tahapan ini, kata-kata yang sebenarnya merujuk pada konteks yang sama akan terkumpul menjadi satu bentuk, sehingga pola sentimen lebih mudah ditangkap oleh model. Contohnya pada baris ke -7, istilah “kkn” diperluas menjadi “korupsi kolusi dan nepotisme”, perubahan

yang dilakukan mengurangi ambiguitas dan memperjelas konteks makna, serta representasi fitur lebih informatif.

Tabel 4.5 Hasil Tahapan Tokenizing

No	*Tokenizing*
1	[layak, hukuman, mati]
2	[baru, ini, korupsi, yang, dampaknya, aku, ngerasain, bang]
3	[negeri, oplosan]
4	[kok, ganti, kanal]
5	[hukum, mati, bagi, maling, pertamina]
6	[mereka, akan, lolos, kenapa, uangnya, udah, banyak]
7	[korupsi, lah, selagi, ada, kesempatankalau, ketahuan, hukumannya, ringan, kok, hidup, korupsi, kolusi, nepotisme, taik, kucing, hukum, dinegara, konoha]
8	[kalau, mobil, pribadi, isi, bbm, subsidi, apakah, merugikan, negara]
9	[udah, males, bang, liat, konten, lu, lebih, cari, aman, karena, udah, jd, jongos, si, botak, gak, sevokal, dulu]
10	[keren, banget, baru, nemu, ini, chanel]

Pada Tabel 4.5 ditunjukkan tahap *tokenizing* yang dilakukan setelah normalisasi. Tahapan ini mengubah teks yang semula berbentuk kalimat menjadi pecahan token-token kata terpisah. Dengan menerapkan tahapan ini, dapat membantu memastikan bahwa setiap kata yang relevan terhadap sentimen muncul sebagai fitur yang jelas, tidak tercampur dengan tanda baca atau inkonsistensi penulisan. Pada penelitian ini, *tokenizing* dilakukan dengan memastikan konsistensi hasil sebelumnya. Penggunaan tanda baca yang tidak menambah makna sentimen, seperti titik koma, tanda tanya, dan tanda seru, dihapus agar kemungkinan terbentuk token terpisah. Konsistensi penggunaan huruf kecil agar menyeragamkan token. Selain itu, kata singkatan yang relevan dibiarkan membuat token kata sendiri. Contohnya, kalimat “layak hukuman mati” diubah menjadi ['layak', 'hukuman', 'mati'], sehingga setiap kata dapat diproses lebih lanjut pada tahap ekstraksi fitur dan pembelajaran model klasifikasi.

Tabel 4.6 Hasil Tahapan Stopwords Removal

No	*Stopwords Removal*
1	[layak, hukuman, mati]
2	[korupsi, dampaknya, ngerasain, bang]
3	[negeri, oplosan]
4	[ganti, kanal]
5	[hukum, mati, maling, pertamina]
6	[lolos, uangnya]
7	[korupsi, kesempatankalau, ketahuan, hukumannya, ringan, hidup, korupsi, kolusi, nepotisme, taik, kucing, hukum, dinegara, konoha]
8	[mobil, pribadi, isi, bbm, subsidi, merugikan, negara]
9	[males, bang, liat, konten, cari, aman, karena, jongos, botak, sevokal]
10	[keren, nemu, chanel]

Pada tabel 4.6, setelah tahap pemisahan kalimat menjadi per token kata, dilakukan tahap *stopword removal*. Tahapan ini bertujuan menghapus kata-kata umum yang sering muncul, tetapi biasanya kata yang dihapus tidak memuat informasi tambahan pada sentimen, seperti kata hubung atau kata perintah. Tahapan penghapusan *stopword* ini membantu menekan *noise*, mengurangi jumlah fitur, dan membuat proses ekstraksi data berfokus pada kata yang membawa makna opini. Pada penelitian saat ini, *stopword* dilakukan menggunakan bahasa pemograman *python* menggunakan daftar *stopword* bahasa Indonesia yang tersedia di *library NLTK*, serta beberapa *stopword* tambahan. Contoh penerapanya, pada kalimat ['mereka', 'akan', 'lolos', 'kenapa', 'uangnya', 'udah', 'banyak'] disederhanakan menjadi ['lolos', 'uangnya'] karena kata-kata seperti “mereka”, “akan”, dan “banyak” tidak memberi kontribusi informasi tambahan terhadap kelas sentimen pada konteks ini. Dengan menerapkan tahapan ini, dapat memastikan data akan mempertahankan token-token kata yang relevan terhadap sentimen, yang nantinya akan memudahkan tahap ekstraksi data dalam melakukan pembobotan pada kata.

Tabel 4.4 Hasil Tahapan Stemming

No	*Stemming*
1	[layak, hukum, mati]
2	[korupsi, dampak, ngerasain, bang]
3	[negeri, oplos]
4	[ganti, kanal]
5	[hukum, mati, maling, pertamina]
6	[lolos, uang]
7	[korupsi, sempat, tahu, hukum, ringan, hidup, kkn, taik, kucing, hukum, negara, konoha]
8	[mobil, pribadi, isi, bbm, subsidi, rugi, negara]
9	[males, bang, liat, konten, cari, aman, jongos, botak, vokal]
10	[keren, nemu, chanel]

Pada tabel 4.7 disajikan tahap *stemming* yang bertujuan untuk mengubah teks menjadi bentuk dasarnya. Tahapan ini menggunakan bantuan *library NLTK* yang di terapkan melalui bahasa pemograman *python*, agar variasi bentuk kata yang sama tidak dihitung sebagai token yang berbeda. Tujuannya adalah menyatukan kata berimbuhan ke bentuk dasarnya, sehingga pembobotan pada tahap ekstraksi data lebih fokus pada inti makna teksnya, bukan pada variasi imbuhan. Tahapan *stemming* dapat membantu mengurangi dimensi fitur dan membuat pola kata kunci lebih mudah ditangkap oleh model. Contohnya, pada kata “hukuman” disederhanakan menjadi “hukum”, sehingga semua kemunculan kata serupa dapat terkumpul dalam satu fitur yang sama. Dampaknya, representasi teks menjadi konsisten, model dapat belajar lebih stabil, dan kemungkinan model mengenali pola kata kunci dari sentimen sangat kuat karena kata-kata tidak terpecah menjadi beberapa fitur kata. Tahap *stemming* juga mengakhiri rangkaian *preprocessing*, yang selanjutnya data sudah bisa di proses di tahap ekstraksi kata dan permodelan.

4.1.4 Hasil Pembobotan TF-IDF

Setelah proses *preprocessing* selesai, tahap selanjutnya melakukan pembobotan kata menggunakan TF-IDF (*Term Frequency-Inverse Document Frequency*), proses ini menggunakan ekstensi *TfidfVectorizer* dari *library sckit-learn*. Tahapan ini bertujuan memberikan nilai bobot pada setiap fitur kata berdasarkan tingkat kepentingannya pada masing-masing dokumen, sehingga kata yang mengandung informasi relevan mendapatkan bobot yang lebih tinggi. Hasil dari pembobotan TF-IDF menghasilkan nilai frekuensi kata terbobot sebanyak 6447, dari 4527 data. Jumlah fitur menunjukkan setiap komentar memiliki karakteristik kata setelah proses pembersihan data. Hasil data pembobotan ini sangat berdampak pada proses klasifikasi menggunakan algoritam *Random Forest*. Jadi semakin baik proses ini, semakin besar peluang model mengenali pola masing-masing sentimen. Contoh pembobotan kata TF-IDF sebagai berikut:

	Dokumen	Kata	Score pembobotan
0	Dokumen 1 (Positif)	layak hukum	0.6667
1	Dokumen 1 (Positif)	layak	0.5291
2	Dokumen 1 (Positif)	hukum mati	0.3300
3	Dokumen 1 (Positif)	mati	0.3134
4	Dokumen 1 (Positif)	hukum	0.2615
...	...	...	...
65	Dokumen 9 (Negatif)	liat	0.2494
66	Dokumen 9 (Negatif)	bang	0.1617
67	Dokumen 10 (Positif)	nemu	0.6297
68	Dokumen 10 (Positif)	chanel	0.5918
69	Dokumen 10 (Positif)	keren	0.5032

70 rows × 3 columns

Gambar 4.3 Hasil Pembobotan TF-IDF

(*Sumber: Hasil Olah Data*)

Pada gambar 4.3 menampilkan tiga kolom utama dokumen, kata, dan skor pembobotan yang menggambarkan hasil pembobotan kata dari TF-IDF pada subset 10 dokumen yang telah melalui tahap *preprocessing text*. Pada Dokumen 1 (Positif), frasa “layak hukum” memperoleh skor pembobotan tertinggi sebesar 0,6667, diikuti oleh kata “layak” (0,5291)

dan “hukum mati” (0,3300). Skor TF-IDF yang tinggi pada term-term tersebut mengindikasikan bahwa kata-kata ini merupakan representasi paling penting dan karakteristik dari isi dokumen tersebut. Selanjutnya, hasil vektorisasi TF-IDF ini digunakan sebagai fitur pelatihan untuk model Random Forest.

4.1.5 Hasil Pembagian Data

Setelah data komentar berhasil melewati tahap *preprocessing text* dan pembobotan TF-IDF, selanjutnya dilakukan pembagian dataset menjadi data latih (*training set*) dan data uji (*testing set*). Pembagian ini dilakukan menggunakan bahasa pemograman *python* menggunakan fungsi *train test split* dari *library sckit-learn* dengan dua rasio pembagian, yaitu 80:20 dan 70:30. Tujuannya untuk mencegah *overfitting* serta mengukur kemampuan generalisasi model terhadap data baru yang belum dilihat oleh model saat pelatihan.

Tabel 4.5 Distribusi Label Sentimen setelah Train Test Split

Rasio	Set	Positif	Netral	Negatif	Total
80:20	**Train**	2.736	540	345	3.621
	Test	684	135	87	906
70:30	**Train**	2.393	473	302	3.168
	Test	1.027	202	130	1.359

Pada tabel 4.7 menampilkan distribusi label setelah dataset dibagi dalam tahapan train-test split menjadi dua rasio, yaitu 80:20 dan 70:30. Pada rasio 80:20, data latih terdiri dari 2.736 sampel berlabel positif, 540 sampel berlabel netral, dan 345 sampel berlabel negatif, dengan total 3.621 sampel. Sementara itu, data uji terdiri dari 684 sampel positif, 135 sampel netral, dan 87 sampel negatif, dengan total 906 sampel. Sedangkan pada rasio 70:30, data latih terdiri dari 2.393 sampel positif, 473 sampel netral, dan 302 sampel negatif, dengan total 3.168 sampel, sedangkan data uji terdiri dari 1.027 sampel positif, 202 sampel netral, dan 130 sampel negatif, dengan total 1.359 sampel.

4.1.6 Penerapan Klasifikasi Random Forest

Setelah data melalui tahapan *preprocessing* dan pembobotan TF-IDF, penelitian ini melakukan klasifikasi sentimen menggunakan algoritma Random Forest. Sub bab ini membahas pembentukan model pada data latih, kemudian menguji kinerja pada data uji. Evaluasi hasil klasifikasi disajikan menggunakan *confusion matrix* dan metriks performa untuk melihat kemampuan model pada setiap label sentimen. Selain itu, sub bab ini juga menyajikan visualisasi *wordcloud* untuk mengidentifikasi kata yang dominan muncul pada seluruh data komentar maupun pada masing-masing label sentimen.

4.1.6.1 Hasil Pembentukan Model

Pada tahapan ini dilakukan pembenukan model klasifikasi sentimen menggunakan algoritma Random Forest dari *library sckit-learn*. Model ini dibentuk dengan data latih yang sudah melalui tahapan *preprocessing* dan pembobotan kata menggunakan metode TF-IDF. Algoritma Random Forest sendiri merupakan metode *ensemble learning* yang menggabungkan banyak pohon keputusan (*decision tree*). Setiap pohon dilatih secara mandiri, kemudian hasil dari pelatihan tersebut akan dikumpulkan untuk memilih hasil melalui *majority voting* atau pengambilan suara terbanyak dalam menentukan hasil akhir yang lebih stabil dan tahan terhadap *overfitting* dibandingkan dengan pohon keputusan tanggal.

Sebelum proses pelatihan, dataset untuk melatih model dibagi menjadi data latih dan data uji, dengan rasio 80:20 dan 70:30. Pada proses pembentukan model yang menggunakan kelas *RandomForestClassifier* dengan pengaturan *hyperparameter* sebagai berikut. Algoritma Random Forest dikonfigurasi dengan jumlah pohon keputusan yang digunakan sebanyak 800 pohon (n_estimators = 800). Jumlah pohon yang banyak ini bertujuan untuk meningkatkan

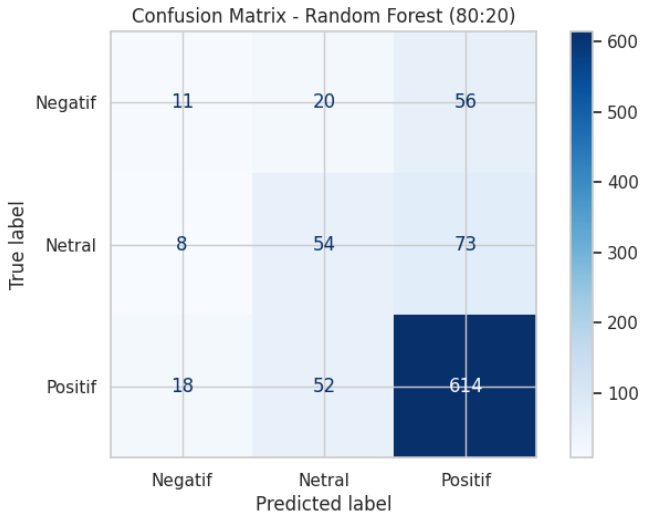
kestabilan prediksi dan mengurangi varians melalui mekanisme *majority voting* dari seluruh pohon. Setiap pohon yang dibentuk menggunakan kriteria pemisahan *node* berbasis *entopy* (criterion =’entropy’), yang membantu model memilih pemisahan data yang paling efektif mengurangi ketidakpastian kelas di setiap *node* percabangan. Untuk menjaga keragaman antar pohon, setiap proses *splitting* hanya dipertimbangkan maksimal log_2 (max_features =’ log_2’) agar setiap pohon keputusan memiliki sudut pandang beragam sekaligus mempercepat proses pelatihan dan mengurangi resiko *overfitting*. Pohon-pohon keputusan dibiarkan tumbuh bebas tanpa batas kedalaman (max_depth= ’None’), sehingga model mampu menangkap hubungan dan pola sentimen yang kompleks dari teks komentar.

Meningat kelas positif mendominasi distribusi label sentimen, diterapkan pengaturan (class_weight =’balanced_subsample’). Parameter ini secara otomatis menyesuaikan bobot setiap kelas pada tiap *subsample bootstrap*, sehingga kontribusi kelas minoritas seperti kelas netral dan negatif meningkat untuk mencegah model bias terhadap kelas mayoritas. Untuk mengurangi kompleksitas model, diterapkan *pruning* dengan parameter (ccp_alpha =’0.001’). Fungsinya untuk memangkas cabang-cabang pohon yang memberikan kontribusi kecil terhadap akurasi, sehingga menghasilkan model yang lebih sederhana. Proses pelatihan ini juga menerapkan teknik *bagging* (bootstrap =’True’) disertai perhitungan *Out-of-Bag score* (oob_score =’True’). Pendekatan ini memungkinkan evaluasi performa model secara internal menggunakan sampel data yang tidak digunakan dalam pelatihan pohon keputusan. Parameter (random_state =’42’), digunakan agar hasil model selalu sama setiap kali dijalankan (*reproduktabilitas*). Sedangkan parameter (n_jobs =’-1’), digunakan untuk memanfaatkan semua *core processor* yang tersedia, sehingga pelatihan 800 pohon keputusan dapat dilakukan secara paralel dan waktu komputasi lebih

efisien. Setelah seluruh pohon keputusan terbentuk, prediksi akhir untuk setiap komentar dihasilkan melalui mekanisme majority voting, yaitu label yang paling banyak diprediksi oleh 800 pohon keputusan akan menjadi hasil klasifikasi sentimen akhir.

4.1.6.2 Evaluasi Kinerja Model

Pada tahapan evaluasi model, dilakukan pengujian menggunakan dua rasio pembagian data, yaitu rasio 80:20 dan 70:30. Evaluasi ini bertujuan untuk mengukur kemampuan model Random Forest dalam mengklasifikasikan label sentimen komentar ke dalam tiga kategori, yaitu label positif, netral, dan negatif. Untuk memudahkan intepretasi, hasil prediksi pada masing-masing skenario kemudian divisualisasikan menggunakan *confusion matrix*. Visualisasi untuk skenario *train test split* rasio 80:20 dan 70:30 sebagai berikut:

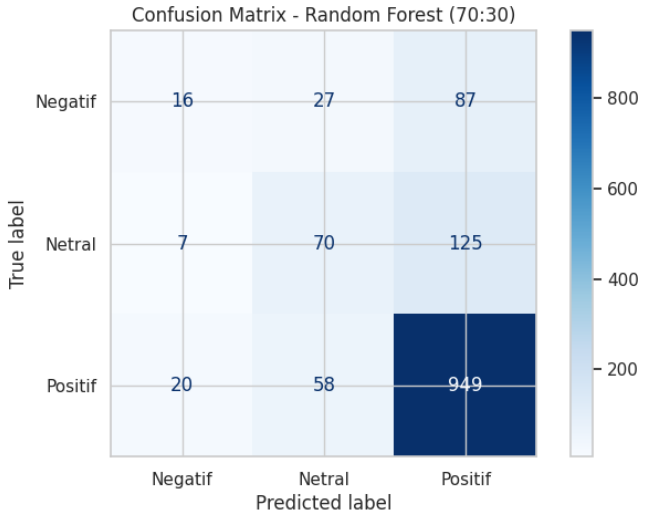


Gambar 4.4 Visualisasi Confusion Matrix Rasio 80:20

(*Sumber: Hasil Olah Data*)

Pada gambar 4.4 disajikan visualisasi *confusion matrix* hasil prediksi model Random Forest pada rasio 80:20. Matriks diatas memperlihatkan perbandingan antara label sebenernya (*true label*) pada

sumbu vertikal dengan label prediksi (*predicted label*) pada sumbu horizontal untuk ketiga kelas sentimen, positif, netral, dan negatif. Dari matriks tersebut dapat dilihat model berhasil memprediksi kelas positif sangat baik, di mana 614 dari 684 komentar label positif berhasil diklasifikasi dengan benar (*true positive*). Namun, sebanyak 52 komentar positif menjadi label netral dan 18 komentar positif menjadi label negatif (*false positive*). Pada kelas netral, dari total 135 komentar, hanya 54 yang diprediksi benar (*true netral*), sementara sisanya 73 komentar diprediksi sebagai positif dan 8 komentar menjadi negatif (*false netral*). Sementara, pada kelas negatif dengan total 89 komentar, sebanyak 11 komentar negatif berhasil diprediksi benar (*true negative*), sementara 20 komentar negatif menjadi label netral dan 56 komentar menjadi label positif (*false negative*).



Gambar 4.5 Visualisasi Confusion Matrix Rasio 70:30

(*Sumber: Hasil Olah Data*)

Pada gambar 4.5 disajikan visualisasi *confusion matrix* hasil prediksi model Random Forest pada rasio 70:30. Sama seperti pada rasio sebelumnya, matriks diatas memperlihatkan perbandingan antara label sebenarnya dan label prediksi. Dari matriks diatas dapat dilihat bahwa model berhasil memprediksi kelas positif sangat baik, karena

label positif mendominasi dataset. Sebanyak 949 dari 1.027 komentar berhasil diprediksi dengan benar (*true positive*), tetapi 58 komentar justru diprediksi label netral dan 20 komentar menjadi label negatif (*false positive*). Pada kelas netral, model berhasil memprediksi dengan benar (*true netral*) sebanyak 70 dari 202 komentar, sementara itu 125 komentar diprediksi menjadi label positif dan 7 komentar menjadi label negatif (*false netral*). Sementara, pada kelas negatif dengan total 130 komentar, sebanyak 16 komentar negatif berhasil diprediksi benar (*true negative*), sementara 27 komentar negatif menjadi label netral dan 87 komentar menjadi label positif (*false negative*). Hal ini merupakan konsekuensi langsung dari ketidakseimbangan distribusi label yang mengakibatkan model lemah dalam memprediksi label minoritas.

Tabel 4.6 Hasil Tabel Evaluasi Model

Rasio	***Accuracy***	***Precision_macro***	***Recall_macro***	***F1_macro***
80:20	0.74	0.51	0.47	0.48
70:30	0.76	0.54	0.46	0.48

Pada tabel 4.6 disajikan hasil evaluasi kinerja model Random Forest menggunakan metrik *macro average* karena distribusi label pada dataset tidak seimbang. Metriks *macro* menghitung rata-rata performa dari setiap kelas dengan bobot yang sama tanpa membedakan kelas yang mendominasi atau tidak (Bagli et al., 2020). Sehingga hasilnya lebih objektif dalam menggambarkan kemampuan model mengenali kelas minoritas dibandingkan melihat dari metrik akurasi yang cenderung bias terhadap kelas mayoritas. Berdasarkan dua skenario pembagian data latih dan data uji. Pada rasio 80:20, model memperoleh nilai *accuracy* sebesar 74%, dengan *precision_macro* 51%, *recall_macro* 47%, dan *f1_macro* 48%. Sementara itu, pada rasio 70:30, model memperoleh nilai *accuracy* sebesar 76%, dengan *precision_macro* 54%, *recall_macro* 46%, dan *f1_macro* 48.

Secara keseluruhan, hasil dari tabel *confusion matrix* menunjukkan kemampuan model dalam memprediksi cukup baik terlihat dari nilai akurasi. Namun, dari metrik *precision*, *recall*, dan *f1-score* masih berada dalam kategori rendah. Hal ini menandakan bahwa model cenderung mampu memprediksi label mayoritas dengan sangat baik, tetapi memiliki keterbatasan dalam memprediksi label minoritas. Sebagai contoh, pada gambar 4.5 dan 4.6, terlihat sebagian komentar negatif salah diprediksi dengan jumlah yang banyak sebagai label positif atau netral. Fenomena ini disebabkan oleh distribusi label data yang sangat tidak seimbang, di mana jumlah komentar negatif jauh lebih sedikit dibandingkan kelas netral dan positif. Akibatnya, model cenderung lebih banyak menebak kelas mayoritas karena pola data pada kelas tersebut jauh lebih kaya dan dominan, sehingga pola sentimen negatif yang lebih sedikit tidak cukup dipelajari dengan baik.

Tabel 4.7 Evaluasi per Kelas Rasio 80:20

Kelas	Precision	Recall	F1-Score	Support
Negatif	0.2973	0.1264	0.1774	87
Netral	0.4286	0.4000	0.4138	135
Positif	0.8264	0.8977	0.8605	684
Macro avg	0.5174	0.4747	0.4839	906
Weighted avg	0.7163	0.7494	0.7284	906

Pada tabel 4.7 disajikan hasil evaluasi kinerja model Random Forest per kelas sentimen. Tabel ini memberikan gambaran yang lebih mendalam, karena menampilkan performa model berdasarkan kelas beserta nilai *support* (jumlah sampel). Model menunjukkan performa yang sangat baik pada kelas positif dengan nilai metrik *precision* sebesar 0,8264, dengan *recall* 0,8977, dan *f1-score* 0,8605 dari total 684 sampel data. Hal ini menunjukkan model mampu mengenali dan mengidentifikasi komentar positif, seperti dukungan terhadap narasi

anti korupsi, dorongan penindakan hukum, serta apresiasi terhadap pembuat konten.

Pada kelas netral, performa model berada pada kategori sedang dengan nilai metrik *precision* 0,4286, *recall* 0,4000, dan *f1-score* 0,4138 dari 135 sampel data. Meskipun kelas netral merupakan salah satu kelas mayoritas, tetapi model masih mengalami kesulitan dalam mengidentifikasi komentar yang bersifat informatif, pertanyaan, candaan, atau kondisi netral terhadap pembahasan. Sementara itu, pada kelas negatif performa model dinilai paling rendah dengan nilai metrik *precision* 0,2973, *recall* 0,1264, dan *f1-score* 0,1774 dari 87 sampel data. Kelas ini merupakan minoritas dari kelas sebelumnya, yang mengakibatkan model hanya berhasil mengidentifikasi lebih sedikit dari label negatif sebenarnya. Sebagian besar komentar negatif berisi pembelaan pelaku korupsi, serangan terhadap pembuat konten, atau penolakan penindakan hukum yang salah diprediksi sebagai positif atau netral.

Tabel 4.8 Evaluasi per Kelas Rasio 70:30

Kelas	Precision	Recall	F1-Score	Support
Negatif	0.3721	0.1231	0.1850	130
Netral	0.4516	0.3465	0.3922	202
Positif	0.8174	0.9241	0.8675	1027
Macro avg	0.5470	0.4646	0.4815	1359
Weighted avg	0.7204	0.7616	0.7315	1359

Pada Tabel 4.8 disajikan hasil evaluasi kinerja model Random Forest per kelas sentimen pada rasio 70:30. Model menunjukkan performa sangat baik pada kelas positif dengan *precision* 0,8174, *recall* 0,9241, dan *f1-score* 0,8675 dari 1.027 sampel. Hal ini membuktikan model sangat mampu mengenali komentar positif yang mendukung narasi anti-korupsi dan apresiasi terhadap konten. Pada kelas netral, performa berada pada kategori sedang dengan *precision* 0,4516, *recall*

0,3465, dan *f1-score* 0,3922 dari 202 sampel. Model masih kesulitan membedakan komentar informatif atau netral terhadap isu korupsi. Sementara pada kelas negatif, performa paling rendah dengan *precision* 0,3721, *recall* hanya 0,1231, dan *f1-score* 0,1850 dari 130 sampel. Model hanya berhasil mendeteksi sebagian kecil komentar negatif yang sebenarnya, sehingga mayoritas salah diprediksi sebagai positif atau netral.

4.1.7 Hasil Visualisasi WordCloud

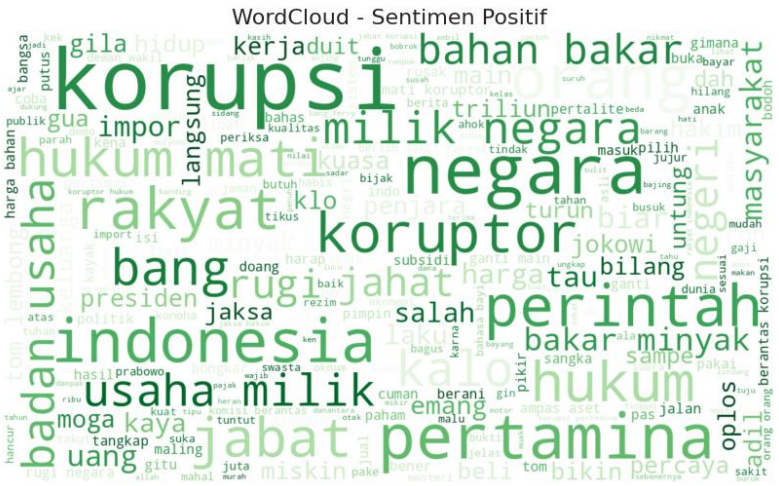
Pada Penelitian ini tidak hanya menyajikan matriks evaluasi kinerja model, tetapi juga menampilkan visualisasi *wordcloud* untuk melihat pola kata yang muncul dalam komentar penonton dari dua video bertopik korupsi di kanal YouTube Malaka Project. Visualisasi ini dibuat menggunakan *library wordcloud* pada bahasa pemograman *python*, di mana ukuran kata yang ditampilkan mencerminkan frekuensi kemunculan katanya dalam dataset. Semakin besar ukuran kata yang ditampilkan, menandakan semakin sering kemunculan kata tersebut.

Gambar 4.6 Visualisasi Wordcloud Semua Komentar

(*Sumber: Hasil Olah Data*)

Pada gambar 4.6 menampilkan visualisasi *wordcloud* untuk semua komentar, terlihat kata "korupsi" sebagai kata paling dominan dengan

ukuran terbesar, diikuti “pertamina”, “negara”, “Indonesia”, “hukum”, “rakyat”, “pejabat”, “uang”, dan “perintah”. Dominasi kata-kata ini menggambarkan fokus utama penonton pada isu korupsi yang terjadi di salah satu BUMN, khususnya Pertamina yang terlibat kasus mega korupsi mengakibatkan kerugian negara sangat besar, serta penyalahgunaan wewenang pejabat publik. Kemunculan kata “negara” dan “rakyat” juga mengindikasikan bahwa kasus korupsi bukan sekedar tindakan individu, tetapi juga pengkhianatan terhadap negara dan merugikan rakyat. Sedangkan, kata “hukum”, “adil”, dan “penjara” yang muncul menunjukkan tuntutan kuat terhadap penegakan hukum tegas dan keadilan. Semntara kata “pejabat” dan “uang” menandakan adanya penyalahgunaan kekuasaan dan terjadi monopoli ekonomi. Secara keseluruhan, pola kata yang muncul pada *wordcloud* memperkuat bahwa video yang disajikan oleh Malaka Project berhasil memicu diskusi publik yang kritis terhadap kasus korupsi sebagai masalah sistematis di Indonesia. Temuan ini selaras dengan misi dari Malaka Project untuk meningkatkan literasi dan pemikiran kritis masyarakat.

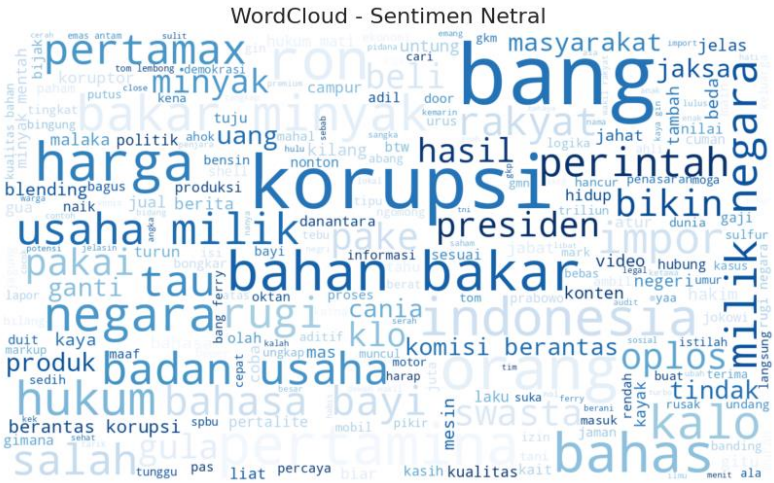


Gambar 4.7 Visualisasi Wordcloud Sentimen Positif

(*Sumber: Hasil Olah Data*)

Pada gambar 4.7 menampilkan visualisasi *wordcloud* untuk label sentimen positif. Terlihat kata “korupsi” paling dominan, diikuti oleh

“hukum mati”, “negara”, “rakyat”, “pertamina”, “adil”, “penjara”, “usaha milik negara”, dan “koruptor”. Pola kata ini mencerminkan dukungan kuat dari penonton terhadap narasi anti korupsi. Dominasi kata “hukum mati” dan “adil” menunjukkan tuntutan penonton dalam menghukum pelaku korupsi dengan tegas. Sementara “pertamina” dan “badan usaha milik negara” mempertegas fokus kasus mega korupsi yang melibatkan salah satu BUMN yang sangat merugikan negara. Sementara itu, kata “jaksa”, “impor”, dan “tom”, menandakan tidak setujunya masyarakat terhadap dakwaan jaksa kepada Tom Lembong mengenai kasus impor gula. Secara keseluruhan, *wordcloud* sentimen positif menggambarkan bahwa video yang disajikan Malaka Project berhasil membangkitkan kesadaran dan dukungan publik untuk pemberantasan korupsi yang lebih tegas.



Gambar 4.8 Visualisasi Wordcloud Sentimen Netral

(*Sumber: Hasil Olah Data*)

Pada gambar 4.8 menampilkan hasil visualisasi *wordcloud* untuk label sentimen netral. Terlihat kata “korupsi” masih dominan, diikuti “harga”, “pertamax”, “minyak”, “pertamina”, “negara”, “bahan bakar”, “badan usaha milik negara”, dan “Indonesia”. Pola kata ini mencerminkan komentar netral yang bersifat informatif, berbasis fakta, dan observasi, di mana penonton lebih banyak menyoroti dampak ekonomi secara langsung terhadap harga BBM dan biaya hidup masyarakat tanpa mengindikasikan

penolakan atau dukungan mengenai isu korupsi. Dominasi kata “harga” dan “minyak” mengindikasikan buruknya audit BUMN, sementara “pertamina” dan “badan usaha milik negara” memfokuskan bahwa masalah ini melibatkan institusi negara. Selain itu, kemunculan kata “jaksa” dan “gkm” menyinggung kebijakan impor gula kristal mentah yang dituduh sebagai tindakan korupsi yang dilakukan oleh Tom Lembong. Secara keseluruhan, *wordcloud* sentimen netral menggambarkan penonton menyikapi video yang disajikan Malaka Project secara rasional dan objektif.

Gambar 4.9 Visualisasi Wordcloud Sentimen Negatif

(*Sumber: Hasil Olah Data Peneliti*)

Pada gambar 4.9 menampilkan visualisasi *wordcloud* untuk label sentimen negatif. Terlihat kata “korupsi” tetap paling dominan, diikuti “pertamina”, “bahan bakar”, “minyak”, “pertamax”, “jabat”, “salah”, “rugi”, “negara”, “harga”, dan “swasta”. Pola kata ini mencerminkan komentar yang membela dan menolak pelaku tindak pidana korupsi yang berada pada video, di mana penonton berusaha menyepelekan dan membantah kasus ini merugikan harga BBM. Kata-kata seperti “salah”, “rugi”, “jabat”, dan “swasta” digunakan untuk mengkritik video sebagai hal yang berlebihan, dengan membandingkan kasus korupsi yang dilakukan perusahaan swasta agar kasus korupsi tidak dilihat sebagai

sebuah kerugian besar. Secara keseluruhan, *wordcloud* sentimen negatif menggambarkan beberapa penonton yang menolak sudut pandang dari video Malaka Project dan menyerang narasi video secara berlebihan.

4.1.8 Hasil Temuan Kesalahan Prediksi

Berdasarkan evaluasi kinerja pada sub bab sebelumnya, terlihat model Random Forest memperoleh hasil akurasi 75% pada rasio pembagian data 80:20 dan 76% pada rasio 70:30. Namun, hasil tersebut tidak merepresentasikan kinerja model secara keseluruhan dalam mengenali seluruh kelas sentimen. Jika ditinjau lebih lanjut berdasarkan tabel evaluasi per kelas, kelemahan utama model terletak pada kemampuan mengenali kelas minoritas, khususnya kelas negatif. Pada rasio 80:20 memperoleh nilai *recall* sebesar 0,1264, sedangkan pada rasio 70:30 lebih rendah, yaitu 0,1231. Nilai *recall* rendah yang diperoleh model menunjukkan sebagian besar komentar berlabel negatif (*true negative*) tidak berhasil dikenali oleh model dan banyak diprediksi menjadi kelas lain.

Tabel 4. 9 Contoh Kesalahan Prediksi Pada Model Random Forest

Komentar	Label Aktual	Label Prediksi
Memangnya klo produksi minyak pakai minyak dalam bisa menjamin tidak terjadi korupsi? Memang isinya udah busuk semua	Negatif	Positif
Orang pada ngamok karna banyak korupsi di tahun ini padahal kalo itu dari jaman sby aja gua udah mikir pasti udah banyak pemain di pemerintahan,but yg bikin gua terkaget-kaget justru knapa pas di kepemimpinan pak prabowo banyak yg kena spill,apakah ini ulah pak prabowo dan kawan akupun tak tau	Negatif	Positf
Menurut dugaan abang, kira-kira presiden sama menteri BUMN tahu nggak korupsi ini sedari dulu-dulu?	Netral	Positif

Berdasarkan Tabel 4.9, memperlihatkan perbandingan antara label aktual yang divalidasi oleh ahli bahasa dengan label hasil prediksi oleh

model Random Forest. Pada sampel data tersebut, terlihat bahwa komentar secara aktual berlabel negatif dan netral secara keliru diprediksi oleh model menjadi komentar berlabel positif. Hal ini menunjukkan bahwa model saat ini mengalami bias terhadap label positif, karena ketimpangan distribusi label yang sangat ekstrem menyebabkan model belajar lebih banyak label mayoritas dibanding label minoritas.

4.1.9 Hasil Pengklasifikasian Sentimen terhadap Fikih Perkataan

Hasil analisis menggunakan algoritma *Random Forest* tidak hanya menunjukkan performa model dalam mengenali pola teks, tetapi juga memberikan *insight* mengenai perilaku komunikasi masyarakat di media sosial jika ditinjau dari perspektif nilai-nilai keislaman. Tujuan dari integrasi ini untuk menggambarkan data aktual mengenai kesadaran moral masyarakat menghadapi isu korupsi yang tercermin pada sentimen positif dari kedua video yang membahas isu ini. Menurut pandangan Islam, setiap kata yang terucap maupun tertulis yang bisa diakses oleh masyarakat merupakan representasi dari ekspresi yang harus dipertanggungjawabkan (Nurhasanah & Suherman, 2022). Terdapat enam prinsip komunikasi, yaitu *Qaulan Sadida*, *Qaulan Ma'rufa*, *Qaulan Baligha*, *Qaulan Maysura*, *Qaulan Layyina*, dan *Qaulan Karima*. Hal ini bertujuan untuk melihat persebaran opini di kanal Malaka Project khususnya pada dua video yang membahas isu korupsi mencerminkan prinsip *tabayyun* dan diikuti akhlak dalam berpendapat di media sosial. Berikut adalah tabel relevansi temuan komentar dengan prinsip-prinsip Fikih Perkataan:

Tabel 4. 10 Relevansi Temuan Komentar dengan Prinsip Fikih Perkataan

Jenis *Qaul*	Makna/Prinsip	Representasi dalam Temuan Penelitian	Label Sentimen
Qaulan Sadida	Perkataan yang tulus, jujur, dan valid.	Penonton yang menyampaikan alasan berbasis data mengenai kerugian negara akibat korupsi.	Positif/ Netral

Jenis *Qaul*	Makna/Prinsip	Representasi dalam Temuan Penelitian	Label Sentimen
Qaulan Ma'rufa	Perkataan yang baik, sopan, dan tepat situasi.	Komentar berupa doa dan dukungan bagi pembuat konten serta apresiasi atas konten.	Positif
Qaulan Baligha	Perkataan yang efektif, jelas, dan membekas di jiwa.	Komentar kritik tajam namun tertsruktur, fokus kebijakan impor dan sistem hukum tidak adil.	Positif
Qaulan Maysura	Perkataan yang mudah dipahami dan menenangkan.	Komentar yang menjelaskan kepada sesama penonton dalam membedakan antara korupsi individu dan sistemik.	Netral
Qaulan Layyina	Perkataan yang lembut dan tidak kasar.	Komentar berisi masukin atau pertanyaan kritis kepada tim Malaka yang disampaikan dengan bahasa yang baik.	Netral
Qaulan Karima	Perkataan yang mulia dan menghormati orang lain.	Komentar yang memuji dan mengucapkan kata sapaan dengan maksud menghormati pembuat konten.	Positif

Berdasarkan tabel 4.10, mayoritas sentimen positif yang mencapai 75%-78% dalam penelitian ini mencerminkan adanya penerapan *Qaulan Sadida* dan *Qaulan Ma'rufa,* hal ini tergambar pada kolom komentar yang berisi dukungan penonton dalam menyuarakan kebenaran antikorupsi. Namun, ditemukan pula penggunaan kata-kata kasar seperti kata 'bajingan' atau 'bangsat' yang diberi label positif karena mendukung narasi antikorupsi pada kedua video, meskipun secara etika komunikasi Islam bertentangan dengan prinsip *Qaulan Layyina*. Hal ini menunjukkan bahwa kesadaran moral masyarakat sangat tinggi menyikapi isu korupsi, tetapi penyampaian di ruang digital masih mengabaikan prinsip etika komunikasi islam.

4.2 Pembahasan

Pembahasan pada penelitian ini disusun melalui rangkaian tahapan yang sistematis dan terstruktur. Tahapan dimulai dari pengambilan data komentar pada kanal YouTube Malaka Project menggunakan *Google App Script* hingga terkumpul 4.544 komentar. Selanjutnya, data tersebut dibersihkan dari duplikasi dan kehilangan baris dengan bantuan ekstensi pada bahasa pemograman *python* sebelum masuk ke tahap selanjutnya. Tahap selanjutnya, melakukan pelabelan data yang dilakukan oleh tiga ahli bahasa berdasarkan kriteria pelabelan yang sudah disepakati. Tahapan ini berfungsi sebagai validasi awal untuk panduan mesin mengklasifikasikan sentimen.

Setelah data bebas duplikasi dan diberi label awal, dataset kemudian di proses melalui tahapan *preprocessing text* untuk membersihkan teks dan menyiapkan teks agar lebih terstruktur ketika diolah oleh algoritma. Rangkaian tahapannya meliputi *cleaning* (menghapus *noise* seperti simbol, *url*, dan elemen lainnya), *case folding* (mengubah teks menjadi huruf kecil), serta *normalization* menggunakan kamus slang yang dibuat peneliti, *tokenizing* untuk memecah kalimat menjadi token kata, *stopword removal* untuk menghapus kata umum yang kurang bermakna, dan *stemming* untuk mengubah kata ke bentuk dasar.

Setelah data komentar melalui tahapan *preprocessing text*, dilanjutkan tahapan pembobotan kata dengan metode TF-IDF (*Term Frequency-Inverse Document Frequency*) yang menghasilkan 6.447 fitur kata unik dari total 4.544 dokumen. Proses pembobotan ini menggunakan *TfidfVectorizer* untuk merepresentasikan fitur optimal. Adapun parameter yang digunakan, seperti *n_gram_range=(1,2)* diaktifkan agar model tidak hanya untuk menangkap *unigram* atau kata tanggal, tetapi juga *bigram* atau dua kata, sehingga konteks seperti “korupsi pertamina” atau “hukum mati” dapat terdeteksi. Serta, pengaturan *min_df=2* dan *max_df=0.95* diaktifkan bertujuan untuk menyaring kata yang jarang muncul (muncul kurang dari 2 dokumen) maupun terlalu umum (hampir muncul disemua dokumen), sehingga mengurangi *noise* dan dimensi fitur yang tidak relevan. Selain itu, parameter *subliner_tf=True*

diaktifkan untuk menerapkan skala logaritmik pada *term frequency*, yang mencegah kata yang sering muncul sangat sering pada satu komentar mendominasi bobot berlebihan. Sedangkan, parameter *smooth_idf=True* diaktifkan untuk menghindari nilai nol pada IDF dan menghasilkan bobot lebih stabil.

Selanjutnya dilakukan tahapan pembagian data latih dan data uji (*train test split*), dengan dua skenario, yaitu rasio 80:20 dan 70:30. Pembagian ini merupakan tahapan penting, karena untuk memastikan model tidak hanya mampu mempelajari pola data, tetapi juga bisa bekerja dengan baik pada data baru yang belum pernah dilihat model sebelumnya. Terlihat pada tabel 4.7, yang menunjukkan distribusi label sentimen yang tidak seimbang. Label positif sangat mendominasi, diikuti label netral, sementara label negatif menjadi label minoritas di semua skenario. Ketidakseimbangan label sentimen ini mencerminkan realitas respon penonton terhadap kedua video yang membahas topik korupsi, di mana mayoritas penonton mendukung narasi anti korupsi, walaupun sebagian bersikap netral, dan sebagian kecil menunjukkan penolakkan. Hal ini sangat berpengaruh terhadap evaluasi model, karena model akan cenderung lebih mudah mempelajari kelas mayoritas, sehingga menurunkan kemampuan prediksi pada kelas minoritas.

Pada tahapan pelatihan model, menggunakan hasil dari pembobotan TF-IDF sebelumnya. Data kemudian dibagi menjadi data latih dan data uji melalui dua skenario rasio, yaitu 80:20 dan 70:30. Model Random Forest bekerja dengan membentuk banyak pohon keputusan menggunakan sampel data yang berbeda, kemudian digabungkan semua pohon keputusan melalui mekanisme *majority voting* agar model tidak mengalami bias pada pola tertentu. Dalam mendukung proses pelatihan model agar mendapatkan hasil terbaik, penelitian ini menerapkan pengaturan parameter yang digunakan selama pelatihan model. Parameter tersebut mencakup pengaturan jumlah pohon (*n_estimators*), pemilihan fitur (*max_features*), pembatasan kompleksitas atau kedalaman pohon (*max_depth*), serta penyeimbangan kelas (*class_weight*) sebagai solusi

awal dari masalah *imbalanced class*, serta dukungan internal sebagai kontrol kualitas menggunakan (*bootstrap* dan *oob_score*).

Dengan pembentukan model Random Forest yang telah dioptimalkan menggunakan pengaturan *hyperparameter* terbaik, kemudian model ini diuji performanya dalam mengklasifikasikan sentimen. Tahapan pengujian ini menggunakan *confusion matrix* dengan dua rasio pembagian data, yaitu 80:20 dan 70:30. Terlihat pada tabel 4.6, pada rasio 80:20 model Random Forest memperoleh hasil metrik *accuracy* 0,74, yang berarti secara keseluruhan model memprediksi kelas dengan benar 74% dari seluruh data uji. Namun, karena permasalahan ketidakseimbangan label pada dataset. Pada tabel 4.7 memperlihatkan kelas positif memperoleh hasil terbaik, hal ini menunjukkan model mampu mendeteksi komentar positif dan berhasil mengidentifikasi sebagian besar data positif dengan akurat. Berbanding terbalik dengan kelas netral dan negatif, model menunjukkan performa yang menurun signifikan. Hal ini menandakan model kesulitan mengidentifikasi komentar negatif dan netral. Peneliti menggunakan metrik *macro average*, untuk menampilkan hasil rata-rata dari metrik. Pada metrik *precision_macro* memperoleh hasil *0*,51 menunjukkan ketepatan prediksi model untuk masing-masing kelas masih kurang bagus, terlihat dari banyaknya kesalahan prediksi pada label minoritas dan model mengalami bias pada label mayoritas. Sedangkan pada metrik *recall_macro* memperoleh hasil 0,47 menandakan kemampuan model menemnukan kembali seluruh data pada tiap kelas masih rendah, khususnya pada kelas minoritas. Dampaknya terlihat pada nilai metrik *f1_macro* 0,48 yang dikategorikan rendah, secara keseluruhan performa antar kelas belum stabil meskipun akurasi menunjukkan hasil yang bagus. Sementara itu, nilai metrik *oob_score* 0,73 menunjukkan hasil performa internal model relatif konsisten dan tidak jauh dari hasil akurasi data uji.

Berdasarkan perbandingan kedua skenario pengujian, model dengan rasio pembagian data70:30 ditetapkan sebagai skenario terbaik dalam model ini. Pada rasio 70:30, model memperoleh hasil metrik *accuracy* 76%, sedikit lebih tinggi dibanding rasio sebelumnya. Peningkatan performa ini terjadi karena

pada rasio ini proprosi data latih cukup optimal bagi model, serta proporsi data uji yang lebih besar memberikan evaluasi yang lebih representatif, sehingga model terhindar dari kecenderungan bias pada kelas mayoritas. Meskipun dikategorikan sebagai rasio terbaik secara keseluruhan, tantangan utama terlihat jika ditinjau menggunakan metrik *macro average*. Terlihat pada metrik *precision_macro* memperoleh nilai 0,54 menandakan prediksi rata-rata kelas sedikit membaik meskipun masih lemah terhadap label minoritas. Namun, nilai metrik *recall_macro* 0,46 sedikit menurun, yang menunjukkan model masih kesulitan menemukan kembali semua data dari setiap label, terutama pada label minoritas. Sesuai dengan hasil dari metrik sebelumnya, nilai metrik *f1_macro* 0,48 juga mengalami sedikit penurunan, yang disebabkan dari hasil metrik *precision* dan *recall* yang rendah pada kelas negatif dan netral. Sementara itu, metrik *oob_score* memperoleh nilai 0,73, menunjukkan kemampuan yang konsisten pada performa internal. Jadi peningkatan pada hasil akurasi di rasio ini tidak merepresentasikan peningkatan kualitas klasifikasi yang merata, karena model masih bias kepada label mayoritas. Dimana nilai *macro average* yang relatif rendah menegaskan model sangat baik dalam memprediksi label mayoritas, tapi faktanya prediksi rata-rata tiap kelas masih rendah akibat ketidakseimbangan distribusi label.

Selain meninjau nilai dari akurasi dan metrik evaluasi lainnya, perlu dilihat pola prediksi model pada masing-masing rasio pembagian data yang ditampilkan pada visualisasi *confusion matrix*. Pada gambar 4.4 menyajikan visualisasi pada rasio 80:20, terlihat model memiliki jumlah data latih yang lebih banyak sehingga cenderung mampu mengenali pola sentimen yang sering muncul pada komentar positif, yang tercermin dari banyaknya prediksi benar pada label positif dibanding label lain pada visualisasi tersebut. Sedangkan, pada rasio 70:30 terlihat bahwa label netral dan terutama label negatif mengalami jumlah kesalahan klasifikasi yang jauh lebih tinggi, di mana banyak komentar negatif justru diprediksi sebagai positif atau netral karena jumlah sampelnya sedikit dan polanya kurang terwakili saat pelatihan. Karena dominasi label positif, mengakibatkan model kurang kuat dalam menangani

label minoritas sehingga komentar dengan sentimen yang tepat atau berlawanan bisa tidak terbaca dengan baik oleh model.

Dengan melihat lebih teliti hasil evaluasi per kelas pada tabel 4.7 (rasio 80:20) dan 4.8 (rasio 70:30), menunjukkan kelemahan dari model ini. Di mana model Random Forest sangat handal dalam mengenali pola sentimen positif. Hampir semua nilai *precision*, *recall*, dan *f1-score* nya memperoleh hasil rata-rata yang tinggi pada kedua rasio. Hal ini, sejalan dengan respon penonton yang cenderung mendukung narasi anti korupsi dan mengapresiasi pembuat konten, sehingga model sangat mudah mempelajari pola kata-kata yang berisi sentimen positif. Sedangkan, kemampuan model dalam mengenali sentimen netral termasuk dalam kategori yang sedang, karena masih terdapat kesalahan prediksi yang cukup banyak seperti sentimen netral menjadi label positif atau negatif. Namun, model dinilai sangat buruk dalam mengenali sentimen negatif, di mana terlihat pada *recall* 0,12. Artinya, model hampir tidak mampu menangkap komentar-komentar yang membela pelaku korupsi. Kelemahan ini bukan sepenuhnya disebabkan karena algoritma yang digunakan tidak sesuai atau buruk, melainkan karena distribusi label sentimen negatif jauh lebih sedikit dibandingkan label positif dan netral. Akibatnya, meskipun nilai metrik akurasi cukup bagus, tetapi model seolah buta terhadap label minoritas yang menjadi salah satu bentuk respon masyarakat dalam menyikapi isu korupsi.

Selain faktor ketidakseimbangan label yang ekstrem, kesalahan ini juga dipengaruhi oleh kompleksitas makna komentar yang sulit dipahami secara utuh oleh ekstraksi fitur TF-IDF. Seperti yang terlihat pada sampel data di tabel 4.9, beberapa komentar berlabel negatif dan netral justru diprediksi sebagai label positif. Contoh komentar label negatif ‘Memangnya klo produksi minyak pakai minyak dalam bisa menjamin tidak terjadi korupsi? Memang isinya udah busuk semua’ Konteksnya sudah tidak memiliki rasa kepercayaan terhadap sistem pemerintahan yang berakibat dalam memandang kasus ini hal biasa terjadi. Namun, model memprediksinya sebagai sentimen positif karena mengandung kata kunci dominan seperti 'minyak' dan 'korupsi' yang secara statistik memiliki frekuensi tinggi pada pola data latih kelas mayoritas (positif).

Model cenderung bias oleh kata-kata yang muncul dikelas positif, sehingga terjadi kesalahan prediksi ketika memaknai kalimat secara utuh.

Melihat fenomena tersebut, distribusi sentimen antar kedua video memberikan gambaran yang cukup jelas. Pada video pertama yang berjudul "Korupsi Terstruktur dan Terjahat Pertamina", sentimen positif mendominasi dengan jumlah 2.855 komentar (75%), diikuti sentimen netral 542 komentar (14%), dan sentimen negatif berjumlah 411 komentar (11%). Hal ini menunjukkan penonton cenderung memberikan dukungan terhadap pembuat konten yang menyuarakan narasi anti korupsi, khususnya dalam kasus yang melibatkan BUMN besar yang merugikan negara dalam skala triliunan rupiah. Menurut data yang dirilis, kerugian negara akibat kasus Pertamax oplosan yang dilakukan pertamina mencapai Rp.193,7 triliun utuk tahun 2023, dan jika diurutkan sepanjang periode 2018-2023 diperkirakan hampir mecapai Rp.968,5 triliun (Tempo, 2025).

Sedangkan, pada video kedua yang berjudul "Vonis Tom Lembong Adalah Masalah Publik", sentimen positif masih mendominasi dengan 565 komentar (79%), diikuti sentimen netral 133 komentar (18%), dan sentimen negatif sangat rendah dengan 21 komentar (3%). Meskipun terdapat kejanggalan dalam putusan Pengadilan Negeri Jakarta pada 18 Juli 2025 karena hakim mengatakan bahwa Tom Lembong tidak menerima keuntungan pribadi dan tidak memiliki niat jahat, meskipun tetap dijatuhi hukuman 4 tahun 6 bulan dan denda sebesar 750 juta. Beberapa dakwaan lainnya dianggap keliru di persidangan, seperti mengatakan bahwa gula pasir saat itu sedang surplus, namun sebenarnya yang terjadi sedang mengalami defisit di karenakan kebutuhan gula meningkat saat bulan Ramadhan (Tempo, 2025). Meskipun demikian, mayoritas penonton tetap mengapresiasi upaya Malaka Project dalam menyajikan kasus tersebut.

Temuan distribusi sentimen diperkuat oleh visualisasi *wordcloud*. Kata-kata yang muncul pada visualisasi ini tidak hanya ukuran dari frekuensi, melainkan juga cerminan dari emosi yang tertuang di kolom komentar. Pada sentimen positif, kemunculan kata "korupsi", "hukum mati", "adil", dan

“pertamina” mendominasi dengan ukuran terbesar. Hal ini, mencerminkan tuntutan kuat dari penonton terhadap penegakan hukum yang tegas bagi pelaku. Contoh opini yang mendukung, “Hukum mati bagi maling pertamina”, sementara yang lain mengatakan “Gila banget orang2 Pertamina kemaruk duit ini!”, dan beragam komentar lainnya yang mengecam tindakan korupsi. Kalimat-kalimat ini menunjukkan bahwa penonton tidak hanya marah dan kecewa, tetapi juga menghubungkan kasus ini dengan sistem peradilan yang kurang baik dalam penerapannya.

Pada sentimen netral banyak menampilkan kata “harga”, “pertamax”, “minyak”, dan “BBM”, yang menunjukkan sebagian penonton lebih fokus pada dampak ekonomi langsung dari kasus korupsi yang terjadi. Contoh opini yang mendukung, “tolong dibuat pencerahan kenapa BBM swasta disuruh beli ke pertamina oleh mentri esdm” dan “Kalau mobil pribadi isi bbm subsidi apakah merugikan negara?”. Kalimat-kalimat ini menunjukkan respon memberikan sebuah informasi dan observasional oleh penonton yang mencoba memahami mekanisme dan dampak yang ditimbulkan, tanpa mengecam atau membela secara langsung.

Pada sentimen negatif kemunculan kata “salah”, “swasta”, “rugi”, dan “jabat”, yang menunjukkan terjadi upaya pembelaan secara implisit terhadap pelaku korupsi dengan membandingkan kasus korupsi yang terjadi di BUMN dengan perusahaan swasta. Contoh opini yang mendukung, “Kita masyarakat cari aman saja. Urusan BBM kan bukan monopoli negara, swasta boleh bermain juga. Ya sudah beli saja di SPBU non-Pertamina. Pertamina auto tutup.” dan “Klo apa yang dilakukan pertamina itu dilakukan oleh perusahaan swasta apakah itu korupsi”. Kalimat-kalimat ini mencoba mengecilkan permasalahan dari korupsi yang dilakukan, sehingga pelaku tidak dianggap sepenuhnya salah.

Temuan penelitian ini sejalan dengan beberapa studi terkait analisis sentimen pada *platform* media sosial di Indonesia, seperti pada penelitian tentang opini publik terhadap UU TPKS di Twitter yang mengalami distribusi label tidak seimbang, dengan total hampir 81% di dominasi label negatif dan

19% label positif, dengan akurasi Random Forest mencapai 85,35% menggunakan TF-IDF dan *class weighting* (Mawar & Rahardi, 2025). Jika dibandingkan dengan hasil penelitian ini yang juga menerapkan teknik *imbalance class* seperti *class weighting*, akurasi yang diperoleh hanya mencapai 74%-76%, dengan performa lebih rendah pada kelas minoritas dengan metrik *recall* negatif 0,12-0,14. Hal ini terjadi karena model cenderung memprediksi pada kelas mayoritas, mengingat adanya ketidakseimbangan label yang ekstrem, di mana jumlah komentar negatif jauh lebih sedikit dibandingkan komentar positif.

Sementara itu, pada penelitian serupa yang menerapkan Random Forest pada analisis sentimen komentar tentang Islamofobia yang menggunakan 1.000 sampel data mencapai akurasi tertinggi 79% pada skenario *train-test split* 90:10, dengan melakukan tahapan *preprocessing* dan TF-IDF serupa (Afdhal et al., 2022). Hal ini sejalan dengan hasil pada penelitian ini yang menunjukkan stabilitas akurasi pada kedua skenario *train-test split* berbeda. Pada rasio 80:20 memperoleh 75% dan 70:30 memperoleh 76%, yang menunjukkan kemampuan cukup baik dalam teks bahasa Indonesia. Namun, terdapat perbedaan pada metrik *macro average* yang cenderung lebih rendah pada penelitian ini, yang disebabkan oleh ketidakseimbangan ekstrem pada label.

Hasil penelitian ini tidak hanya membuktikan bahwa algoritma Random Forest cukup baik dalam mengklasifikasikan komentar YouTube. Hal ini terlihat dari tingkat akurasi yang stabil pada masing-masing rasio pembagian data. Selain itu, temuan ini juga memberikan banyak *insight* yang menggambarkan kolom komentar menjadi ruang diskusi digital, sebagai tempat masyarakat saling bertukar pendapat dan berbagi informasi. Dari segi praktis, mayoritas penonton sangat mendukung isi konten yang membahas narasi anti korupsi. Hal ini terlihat dari banyaknya komentar sentimen positif dan beberapa pola kata yang muncul, seperti “hukum mati”, “adil”, serta “pertamina”. Temuan ini bisa membantu tim Malaka Project untuk terus secara konsisten membuat konten mengangkat isu-isu sosial yang dibahas dari sudut pandang edukasi. Hal ini sangat penting dilakukan, karena di tengah banyaknya

kasus korupsi yang terjadi di Indonesia, sebuah konten edukasi bisa menjadi solusi dalam meningkatkan literasi masyarakat melalui *platform* digital, sebelum terdapat banyaknya berita provokatif menyebar yang mempengaruhi pandangan masyarakat.

Selain itu tidak hanya memberikan manfaat secara praktis, temuan ini juga selaras dengan ajaran islam. Sebagaimana yang ditegaskan di dalam Al-Qur'an tentang larangan melakukan segala bentuk korupsi dan mengambil harta orang lain, perintah ini tertuang dalam Q.S. Al-Baqarah ayat 188:

وَلَا تَأْكُلُوْٓا اَمْوَالَكُمْ بَيْنَكُمْ بِالْبَاطِلِ وَتُدْلُوْا بِهَآ اِلَى الْحُكَّامِ لِتَأْكُلُوْا فَرِيْقًا مِّنْ اَمْوَالِ النَّاسِ بِالْاِثْمِ وَاَنْتُمْ تَعْلَمُوْنَ ﴿١٨٨﴾

Artinya: "*Janganlah kamu makan harta di antara kamu dengan jalan yang batil dan (janganlah) kamu membawa urusan harta itu kepada para hakim dengan maksud agar kamu dapat memakan sebagian harta orang lain itu dengan jalan dosa, padahal kamu mengetahui*"

Untuk memperjelas relevansi ayat ini dengan temuan pada kolom komentar yang di dominasi oleh sentimen positif. Hal ini menunjukkan bahwa masyarakat Indonesia memiliki kesadaran moral yang kuat. Ayat tersebut secara tegas melarang segala upaya mendapatkan dan memanipulasi harta melalui cara yang tidak benarkan oleh syariat islam. Menurut tafsir tahlili, pada bagian pertama ayat ini Allah melarang makan harta orang lain dengan jalan yang batil. Konteks "makan" adalah "mempergunakan atau memanfaatkan". Batil sendiri adalah cara yang tidak dianjurkan menurut hukum yang telah ditentukan Allah (Qur'an Kemenag, 2022).

Integrasi temuan ini memberikan pemahaman mendalam pada masyarakat yang merespon dengan menolak tindakan korupsi di negara Indonesia, hal ini selaras dengan nilai tauhid tetapi masih terdapat ketidakselarasan dengan prinsip etika islam dalam menyampaikan pesan di media sosial. Fenomena ini membagi perilaku penonton dalam memandang nilai kejujuran (*Qaulan Sadida*) dalam menjaga hak rakyat lebih penting disbanding menjunjung adab berkomentar di media sosial. Hal ini, menjelaskan kemarahan rakyat atas tindak korupsi sering kali memunculkan diksi keras yang melampaui batas

kelembutan (*Qaulan Layyina*), dengan pemahaman menyuarakan kebenaran dengan kata-kata yang tidak beradab di normalisasikan oleh pengguna media sosial.

Berdasarkan hasil yang diperoleh, temuan ini masih belum maksimal karena terdapat beberapa kelemahan, seperti pada data komentar hanya dikumpulkan pada satu waktu saja, sehingga hasilnya tidak dapat menggambarkan perubahan sentimen masyarakat seiring waktu. Selain itu, analisis hanya mencakup dua video dari kanal Malaka Project, padahal kanal tersebut memiliki banyak konten edukasi lain yang berpotensi memberikan gambaran lebih luas. Proses pelabelan dilakukan secara manual oleh tiga ahli bahasa, sehingga masih mengandung unsur subjektivitas manusia. Keterbatasan paling utama adalah distribusi label yang sangat ekstrem yang di dominasi oleh kelas positif, serta penggunaan hanya satu algoritma Random Forest tanpa perbandingan dengan model lain, sehingga kemampuan model dalam mendeteksi sentimen minoritas masih terbatas. Secara keseluruhan, model Random Forest menunjukkan akurasi yang stabil, sementara distribusi sentimen dan *wordcloud* memberikan gambaran jelas tentang persepsi masyarakat terhadap topik korupsi di kanal Malaka Project.

BAB V
PENUTUP

5.1 Kesimpulan

Berdasarkan hasil analisis sentimen pada dua video yang membahas topik korupsi di kanal YouTube Malaka Project, di mana video pertama berjudul "Korupsi Terstruktur dan Terjahat Pertamina" serta video kedua "Vonis Tom Lembong Adalah Masalah Publik". Dari kedua video tersebut, didapatkan total komentar sebanyak 4.544 komentar, dan data komentar di klasifikasikan menjadi tiga label sentimen (positif, netral, dan negatif) menggunakan algoritma Random Forest dengan pembobotan TF-IDF. Pada rasio pembagian data latih dan data uji 80:20 mencapai tingkat akurasi 74%, sedangkan pada rasio pembagian data latih dan data uji 70:30 mendapatkan tingkat akurasi 76%. Meskipun akurasi keseluruhan cukup stabil, nilai *macro average* metrik *precision*, *recall*, dan *f1-score* masih rendah karena distribusi label yang sangat tidak seimbang, di mana label sentimen positif mendominasi dan model cenderung lemah dalam mendeteksi pola sentimen negatif. Model lebih bias dalam mempelajari data dari label positif menyebabkan kesalahan prediksi sering terjadi pada label minoritas, seperti pada label negatif dan netral.

Selain itu, perbandingan distribusi sentimen antar kedua video menunjukkan pola yang berbeda. Pada video pertama ("Korupsi Terstruktur dan Terjahat Pertamina"), sentimen positif mendominasi dengan 2.855 komentar (sekitar 75 %), diikuti netral (542) dan negatif (411). Pada video kedua ("Vonis Tom Lembong Adalah Masalah Publik"), sentimen positif tetap dominan dengan 565 komentar (sekitar 78,6 %), namun proporsi negatif sangat rendah (hanya 21 komentar). Temuan ini mencerminkan bahwa penonton lebih emosional dan defensif terhadap kasus korupsi individu dibandingkan kasus korupsi sistemik di BUMN. Visualisasi WordCloud semakin memperkuat bahwa masyarakat Indonesia masih memiliki kesadaran moral yang kuat terhadap isu korupsi, sebagaimana selaras dengan larangan mengambil harta yang bukan miliknya dalam QS. Al-Baqarah ayat 188.

5.2 Saran

Berdasarkan kesimpulan dari penelitian ini, terdapat beberapa rekomendasi yang mendukung pengembangan pada hasil penelitian ini menjadi lebih baik lagi, antara lain:

1. Bagi praktisi atau pembuat konten di media sosial, bisa menerapkan tahapan analisis sentimen mulai dari *scraping*, *preprocessing*, hingga visualisasi yang dapat dijadikan *framework* untuk memetakan pola opini publik pada topik korupsi ataupun topik lainnya. Dengan menerapkan proses analisis yang sistematis, hasil identifikasi terhadap kecenderungan respon publik yang tergambar dalam kategori (positif, netral, negatif) dapat menjadi data objektif dalam menyusun narasi konten yang efektif. Sehingga penyampaian pesan tetap relevan terhadap penonton dan sejalan dengan tujuan meningkatkan literasi informasi masyarakat di ruang digital khususnya *platform* YouTube.
2. Bagi penelitian selanjutnya, disarankan untuk melakukan eksplorasi dan komparasi metode penanganan ketidakseimbangan data, seperti teknik *oversampling*, *undersampling*, *class weighting*, atau memilih algoritma yang cocok dalam mengelola data teks komentar di media sosial yang bersifat tidak terstruktur dan kaya akan konteks. Langkah ini penting dalam menangani kasus dominasi terhadap salah satu label, sehingga model gagal mengenali pola di label lainnya.
3. Penelitian selanjutnya diharapkan memilih topik yang memiliki relevansi jangka panjang dan sumber data lebih beragam. Pemilihan topik yang kuat secara isu pembahasan sangat berpengaruh terhadap hasil penelitian, supaya hasilnya tetap valid dan relevan dengan perkembangan informasi yang sangat masif di media sosial.

DAFTAR PUSTAKA

Adek, R. T., Fitri, Z., Siegar, S. C., Informatika, T., Malikussaleh, U., & Machine, S. V. (2025). *Analisis Sentimen Komentar Pada Saluran Youtube Beauty Vlogger Berbahasa Indonesia Menggunakan Metode Support Vector Machine*. *5*(2), 164–175.

Afdhal, I., Kurniawan, R., Iskandar, I., Salambue, R., Budianita, E., & Syafria, F. (2022). Penerapan Algoritma Random Forest Untuk Analisis Sentimen Komentar Di YouTube Tentang Islamofobia. *Jurnal Nasional Komputasi Dan Teknologi Informasi*, *5*(1), 122–130.

Agustina, N., Novita, R., Mustakim, & Rozanda, N. E. (2024). The Implementation of TF-IDF and Word2Vec on Booster Vaccine Sentiment Analysis Using Support Vector Machine Algorithm. *Procedia Computer Science*, *234*, 156–163.

Ahmadi, M. I., Gustian, D., & Sembiring, F. (2021). Analisis Sentiment Masyarakat terhadap Kasus Covid-19 pada Media Sosial Youtube dengan Metode Naive bayes. *Jurnal Sains Komputer & Informatika (J-SAKTI*, *5*(2), 807–814.

Ahmed, N. S., Huynh, N., Gassman, S., Mullen, R., Pierce, C., & Chen, Y. (2022). Predicting Pavement Structural Condition Using Machine Learning Methods. *Sustainability (Switzerland)*, *14*(14).

Alvanof, M. M., Bustami, & Dinata, R. K. (2024). Penerapan Algoritma Random Forest dalam Deteksi dan Klasifikasi Ransomware. *Jurnal Elektronika Dan Teknologi Informasi*, *5*(2), 2721–9380.

Alviyanti, R., & Said, N. H. M. (2025). Respon Publik Terhadap Strategi Dakwah melalui Stand Up Comedy pada Platform YouTube. *Concept: Journal of Social Humanities and Education*, *4*(1), 21–31.

Amin, M. F. (2023). Confusion Matrix in Three-class Classification Problems: A Step-by-Step Tutorial. *Journal of Engineering Research*, *7*(1), 1–10.

Anam, M. K., Fitri, T. A., Agustin, A., Lusiana, L., Firdaus, M. B., & Nurhuda, A. T. (2023). Sentiment Analysis for Online Learning using The Lexicon-Based Method and The Support Vector Machine Algorithm. *ILKOM Jurnal Ilmiah*,

15(2), 290–302.

Andayani, S., & Ryansyah, A. (2017). Implementasi Algoritma TF-IDF Pada Pengukuran Kesamaan Dokumen. *JuSiTik : Jurnal Sistem Dan Teknologi Informasi Komunikasi*, *1*(1), 53.

Anggreny, A. D. (2025). *Sentiment Analysis Of Free Meal Program For School Students Using Algoritma Naive Bayes*. *2*, 484–492.

Ardiani, L., Sujaini, H., & Tursina, T. (2020). Implementasi Sentiment Analysis Tanggapan Masyarakat Terhadap Pembangunan di Kota Pontianak. *Jurnal Sistem Dan Teknologi Informasi (Justin)*, *8*(2), 183.

Arifianto, A. S., Dewi Safitri, K., Agustianto, K., & Wiryawan, I. G. (2022). Pengaruh Prediksi Missing Value pada Klasifikasi Decision Tree C4.5. *Jurnal Teknologi Informasi Dan Ilmu Komputer*, *9*(4), 779–786.

Aufar, A. F., Mochamad Alfan Rosid, Eviyanti, A., & Astutik, I. R. I. (2023). Optimizing Text Preprocessing for Accurate Sentiment Analysis on E-Wallet Reviews. *JICTE (Journal of Information and Computer Technology Education)*, *7*(2), 42–50.

Azzuhri, A. (2020). Tabayyun As a Crucial Aspect in the Quranic Concept of Ummah Analysis of "Tabayyun" in Sura AL-HUJARAAT (49:6). *HUNAFA Jurnal Studia Islamika*, 21.

Bagli, E., Visani, G., & Grandini, M. (2020). *Metrics For Multi-Class Classification: An Overview*. 1–17.

Daraini, Sa'adatut, N., & Masnawati.I, E. (2024). Peran Media Sosial Youtube Sebagai Media Edukasi Dalam Pendidikan Generasi Z. *MIND Jurnal Ilmu Pendidikan Dan Budaya 4.2 (2024)*, *4*(2), 81–87.

Fauzan, A. C., & Hikmah, K. (2022). Implementasi Algoritma Naive Bayes Dalam Analisis Polarisasi Opini Masyarakat Terkait Vaksin Covid-19. *Rabit : Jurnal Teknologi Dan Sistem Informasi Univrab*, *7*(2), 122–128.

Hanani, A. (2023). *Algoritma Term Frequency – Inverse Document Frequency (TF-IDF)*. *December*.

Harianto, F. R. A., Alawi, Z., & Sa'ida, I. A. (2025). Pengaruh Komposisi Split Data Pada Akurasi Klasifikasi Penderita Diabetes Menggunakan Algoritma

Machine Learning. *Jurnal Sistem Informasi Dan Informatika (Simika)*, *8*(1), 36–44.

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.

Himawan, R. D., & Eliyani. (2021). Perbandingan Akurasi Analisis Sentimen Tweet terhadap Pemerintah Provinsi DKI Jakarta di Masa Pandemi. *JEPIN (Jurnal Edukasi Dan Penelitian Informatika)*, *7*(1), 58–63.

Irlanie, C. C. (2025, Juli 25). Vonis Tom Lembong Adalah Masalah Publik.

Irwandi, F. (2025, Maret 4). Korupsi Terstruktur dan Terjahat Pertamina.

Jamil, M., Hadiyanto, H., & Sanjaya, R. (2024). Sentiment Analysis: Classifying Public Comments on YouTube in Disaster Management Simulation in Indonesia Using Naïve Bayes and Support Vector Machine. *Ingenierie Des Systemes d'Information*, *29*(2), 437–446.

Kamil, A. I., Pratiwi, O. N., & Witarsyah, D. (2025). Analisis Sentimen dan Pemodelan Topik terhadap Aplikasi Pembelajaran Online pada Platform Google Play. *JIPI (Jurnal Ilmiah Penelitian Dan Pembelajaran Informatika)*, *10*(2), 836–849.

KPK, K. P. (2024, Desember 19). *Kinerja KPK 2020-2024: Tangani 2.730 Perkara Korupsi, Lima Sektor Jadi Fokus Utama*. Retrieved from KPK-Ruang Informasi Berita: https://www.kpk.go.id/id/ruang-informasi/berita/kinerja-kpk-2020-2024-tangani-2730-perkara-korupsi-lima-sektor-jadi-fokus-utama

Mahfud, F. K. R., Mudawamah, N. S., & Hariyanto, W. (2020). *Sentiment Analysis of Perpustakaan Nasional Republik Indonesia Through Social Media Twitter*. *12*(1), 90–93.

Mario, C., Suryono, R. R., Indonesia, U. T., & Lampung, B. (2025). *Analisis Sentimen Publik pada Film Dirty Vote di Youtube Menggunakan Random Forest dan Naive Bayes*. *10*(1), 111–122.

Mawar, H. S., & Rahardi, M. (2025). *Sentiment Analysis of the TPKS Law on Twitter : A Comparative Study of Classification Algorithm Performance*. *9*(6),

3087–3096.

Muzaki, A., Febriana, V., & Cholifah, W. N. (2024). *Analisis Sentimen pada Ulasan Produk di E-Commerce Dengan Metode Naive Bayes*. *05*(04), 758–765.

Nababan, W. M. C. (2025). Rugikan Negara Rp 193,7 Triliun, Tujuh tersangka Kasus Korupsi Pertamina ditahan. *Kompas*. https://www.kompas.id/artikel/rugikan-negara-rp-1937-triliun-tujuh-tersangka-kasus-korupsi-pertamina-ditahan

Naing, S. Z. S., & Udomwong, P. (2024). Public Opinions on ChatGPT: An Analysis of Reddit Discussions by Using Sentiment Analysis, Topic Modeling, and SWOT Analysis. *Data Intelligence*, *6*(2), 344–374.

Nisrinaf, H. (2023). Kehadiran Malaka Project untuk Mudahkan Akses Pendidikan di Indonesia. *Goodnewsfromindonesia*. https://www.goodnewsfromindonesia.id/2023/10/28/kehadiran-malaka-project-untuk-mudahkan-akses-pendidikan-di-indonesia

Nugroho, M. W. (2025). Analisis Performa Algoritma Random Forest dalam Mengatasi Overfitting pada Model Prediksi. *Jurnal JTIK (Jurnal Teknologi Informasi Dan Komunikasi)*, *9*(4), 1562–1571.

Nurhasanah, N. D. N., & Suherman, Y. N. (2022). *Qaulan Sadida, Qaulan Ma'rufa, Qaulan Baligha, Qaulan Maysura, Qaulan Layyina Dan Qaulan Karima Itu Sebagai Landasan Etika Komunikasi Dalam Dakwah*. *1*(2), 76–84.

Pitaloka, P. S. (2025, Juli 8). *TEMPO*. Retrieved from Tom Lembong Dituntut 7 Tahun Penjara: Kilas Balik Kasus Korupsi Impor Gula: https://www.tempo.co/hukum/tom-lembong-dituntut-7-tahun-penjara-kilas-balik-kasus-korupsi-impor-gula-1935412

Pranata, R. A., Rudiman, & Verdikha, N. A. (2024). Metode Pembobotan TF-IDF Untuk Klasifikasi Teks Quick Count Pemilihan Wakilpresiden Indonesia2024 Pada X Twitter dengan Metode SVM. *Jurnal Teknologi Informasi Jurnal Keilmuan Dan Aplikasi Bidang Teknik Informatika*, *18 No 2*(2), 126–138.

Pratama, M. F. R., & Rivan, M. E. (2024). Analisis Komentar Instagram

Menggunakan Gaussian Naive Bayes. *Scientica*, *3*(7), 2024.

Ramadhani, B., & Suryono, R. R. (2024). Komparasi Algoritma Naïve Bayes dan Logistic Regression Untuk Analisis Sentimen Metaverse. *Jurnal Media Informatika Budidarma*, *8*(2), 714.

Raya, D., Yusnita, A., & Haristyawan, I. (2025). Application of K-Nearest Neighbor Algorithm For Sentiment Analysis On Free Fire Online Game Based On Google Play Store Reviews. *J-Intech*, *13*(01), 96–106.

Salman, H. A., Kalakech, A., & Steiti, A. (2024). Random Forest Algorithm Overview. *Babylonian Journal of Machine Learning*, *2024*, 69–79.

Sathyanarayanan, S. (2024). Confusion Matrix-Based Performance Evaluation Metrics. *African Journal of Biomedical Research*, *27*(4), 4023–4031.

Septiani, D., & Isabela, I. (2022). Analisis Term Frequency Inverse Document Frequency (TF-IDF) Dalam Temu Kembali Informasi Pada Dokumen Teks. *SINTESIA: Jurnal Sistem Dan Teknologi Informasi Indonesia*, *1*(2), 81–88.

Tempo. (2025, July 22). *Kejanggalan vonis Tom Lembong*. Retrieved from Tempo.co: https://www.tempo.co/infografik/infografik/kejanggalan-vonis-tom-lembong-2049528

Tempo. (2025, March 16). *Kerugian Negara Akibat Pertamina Pertamax Oplosan Hampir 1 Kuadriliun, Berikut Bilangan di atas Triliun*. Retrieved from Tempo.co: https://www.tempo.co/sains/kerugian-negara-akibat-pertamina-pertamax-oplosan-hampir-1-kuadriliun-berikut-bilangan-di-atas-triliun-1220215

Utomo, P., & Prayogi, F. (2021). Literasi Digital: Perilaku dan Interaksi Sosial Masyarakat Bengkulu Terhadap Penanaman Nilai-nilai Kebhinekaan Melalui Diseminasi Media Sosial. *Indonesian Journal of Social Science Education (IJSSE)*, *3*(1), 65.

Yuniar, E., Utsalinah, D. S., & Wahyuningsih, D. (2022). Implementasi Scrapping Data Untuk Sentiment Analysis Pengguna Dompet Digital dengan Menggunakan Algoritma Machine Learning. *Jurnal Janitra Informatika Dan Sistem Informasi*, *2*(1), 35–42.

Yusrana, R. D., Josepto, G. B., Aronggear, M. D. R., Pranatawijaya, V. H., &

Priskila, R. (2024). Analisis Sentiment Komentar Video Youtube “Epic Rap Battle of Presidency 2024” Menggunakan Algoritma Naïve Bayes & Svm. *Journal of Engineering and Sustainable Technology*, *10*(02), 1064–1069.

LAMPIRAN

CLEANING TEXT

```
def tahap_cleaning(text):
    if not isinstance(text, str): return ""
    text = html.unescape(text)
    text = re.sub(r'http\S+|www\S+|https\S+', '', text)
    text = re.sub(r'@\w+', '', text)
    text = re.sub(r'#\w+', '', text)
    text = re.sub(r'[^\w\s]', ' ', text)  # Hapus tanda baca
    text = re.sub(r'\d+', ' ', text)      # Hapus angka
    text = re.sub(r'\s+', ' ', text).strip() # Hilangkan spasi berlebih
    return text

df1['cleaning'] = df1['teks'].apply(tahap_cleaning)
print("HASIL TAHAP 1: CLEANING")
display(df1[['teks', 'cleaning']].head())
```

```
HASIL TAHAP 1: CLEANING
```

	teks	cleaning
0	Jangan lupa like, komen, dan share video ini s...	Jangan lupa like komen dan share video ini sup...
1	KORUPTOR BANGSAT	KORUPTOR BANGSAT
2	SETIAP NEGARA PASTI PUNYA OLIGARKINYA MASING M...	SETIAP NEGARA PASTI PUNYA OLIGARKINYA MASING M...
3	JIKA KALIAN INGIN INDONESIA MAJU SILAHKAN BACA...	JIKA KALIAN INGIN INDONESIA MAJU SILAHKAN BACA...
4	<a href="https://www.youtube.com/watch?v=YWBHj...	a href

CASE FOLDING

```
def tahap_case_folding(text):
    return text.lower()

df1['case_folding'] = df1['cleaning'].apply(tahap_case_folding)
print("HASIL TAHAP 2: CASE FOLDING")
display(df1[['cleaning', 'case_folding']].head())
```

```
HASIL TAHAP 2: CASE FOLDING
```

	cleaning	case_folding
0	Jangan lupa like komen dan share video ini sup...	jangan lupa like komen dan share video ini sup...
1	KORUPTOR BANGSAT	koruptor bangsat
2	SETIAP NEGARA PASTI PUNYA OLIGARKINYA MASING M...	setiap negara pasti punya oligarkinya masing m...
3	JIKA KALIAN INGIN INDONESIA MAJU SILAHKAN BACA...	jika kalian ingin indonesia maju silahkan baca...
4	a href	a href

NORMALIZATION

```
def tahap_normalization(text):
    words = text.split()
    normalized = [slang_dict.get(word, word) for word in words]
    return ' '.join(normalized)

df1['normalization'] = df1['case_folding'].apply(tahap_normalization)
print("HASIL TAHAP 3: NORMALIZATION")
display(df1[['case_folding', 'normalization']].head())
```

```
HASIL TAHAP 3: NORMALIZATION
```

	case_folding	normalization
0	jangan lupa like komen dan share video ini sup...	jangan lupa like komen dan share video ini sup...
1	koruptor bangsat	koruptor sialan
2	setiap negara pasti punya oligarkinya masing m...	setiap negara pasti punya oligarkinya masing m...
3	jika kalian ingin indonesia maju silahkan baca...	jika kalian ingin indonesia maju silahkan baca...
4	a href	a href

TOKENIZING

```
def tahap_tokenizing(text):
    return text.split()

df1['tokenizing'] = df1['normalization'].apply(tahap_tokenizing)
print("HASIL TAHAP 4: TOKENIZING")
display(df1[['normalization', 'tokenizing']].head())
```

```
HASIL TAHAP 4: TOKENIZING
```

	normalization	tokenizing
0	jangan lupa like komen dan share video ini sup...	[jangan, lupa, like, komen, dan, share, video,...
1	koruptor sialan	[koruptor, sialan]
2	setiap negara pasti punya oligarkinya masing m...	[setiap, negara, pasti, punya, oligarkinya, ma...
3	jika kalian ingin indonesia maju silahkan baca...	[jika, kalian, ingin, indonesia, maju, silahka...
4	a href	[a, href]

STOPWORD REMOVAL

```
def tahap_stopword_removal(tokens):
    # Filter token: cek apakah ada di stop_words dan panjang karakter > 2
    return [word for word in tokens if word not in stop_words and len(word) > 2]

df1['stopword_removal'] = df1['tokenizing'].apply(tahap_stopword_removal)
print("HASIL TAHAP 5: STOPWORD REMOVAL")
display(df1[['tokenizing', 'stopword_removal']].head())
```

```
HASIL TAHAP 5: STOPWORD REMOVAL
```

	tokenizing	stopword_removal
0	[jangan, lupa, like, komen, dan, share, video,...	[lupa, like, komen, share, video, orang]
1	[koruptor, sialan]	[koruptor, sialan]
2	[setiap, negara, pasti, punya, oligarkinya, ma...	[negara, oligarkinya, bedanya, oligarki, negar...
3	[jika, kalian, ingin, indonesia, maju, silahka...	[indonesia, maju, silahkan, baca, kepala, daer...
4	[a, href]	[]

STEMMING

```
def tahap_stemming(tokens):
    return [stemmer.stem(word) for word in tokens]

print("⏳ Memulai proses stemming (butuh waktu beberapa saat)...")
tqdm.pandas(desc="Progress Stemming")

# Proses Stemming pada list token
df1['stemming_tokens'] = df1['stopword_removal'].progress_apply(tahap_stemming)

# Menggabungkan token kembali menjadi kalimat utuh (string) untuk pemodelan
df1['final_clean_text'] = df1['stemming_tokens'].apply(lambda tokens: ' '.join(tokens))

print("\nHASIL TAHAP 6: STEMMING")
display(df1[['stopword_removal', 'stemming_tokens', 'final_clean_text']].head())
```

```
⏳ Memulai proses stemming (butuh waktu beberapa saat)...
Progress Stemming: 100%|██████████| 4527/4527 [13:16<00:00,  5.69it/s]
HASIL TAHAP 6: STEMMING
```

	stopword_removal	stemming_tokens	final_clean_text
0	[lupa, like, komen, share, video, orang]	[lupa, like, komen, share, video, orang]	lupa like komen share video orang
1	[koruptor, sialan]	[koruptor, sial]	koruptor sial
2	[negara, oligarkinya, bedanya, oligarki, negar...	[negara, oligarki, beda, oligarki, negara, maj...	negara oligarki beda oligarki negara maju indu...
3	[indonesia, maju, silahkan, baca, kepala, daer...	[indonesia, maju, silah, baca, kepala, daerah,...	indonesia maju silah baca kepala daerah presid...
4	[]	[]	

TF-IDF

```
from sklearn.feature_extraction.text import TfidfVectorizer
import pandas as pd

print("🚀 Menyiapkan TF-IDF Vectorizer...")

# Inisialisasi TF-IDF dengan parameter optimal
tfidf = TfidfVectorizer(
    ngram_range=(1, 2),          # Mengambil kata tunggal (Unigram) dan kombinasi 2 kata (Bigram)
    min_df=2,                    # Kata harus muncul di minimal 2 dokumen/baris komentar
    max_df=0.95,                 # Abaikan kata yang muncul di >95% komentar (terlalu umum/noise)
    sublinear_tf=True,           # Skala logaritmik agar kata yang muncul berkali-kali tidak mendominasi
    smooth_idf=True
)

print("✅ Section 1 Selesai: Vectorizer siap digunakan!")
```

```
🚀 Menyiapkan TF-IDF Vectorizer...
✅ Section 1 Selesai: Vectorizer siap digunakan!
```

TRAIN TEST SPLIT (80:20)

```
# ================== VERSI 1: RASIO 80:20 ==================
X_train_80, X_test_20, y_train_80, y_test_20 = train_test_split(
    x_tfidf,                  # Data fitur TF-IDF
    y,                        # Data target label
    test_size=0.2,            # 80% train, 20% test
    stratify=y,               # Menjaga proporsi label agar seimbang antara train & test
    random_state=42           # Mengunci pengacakan agar hasilnya konsisten saat diulang
)

print("📊 SKENARIO 1: RASIO 80:20")
print(f"X_train shape : {X_train_80.shape}")
print(f"X_test shape  : {X_test_20.shape}")
print(f"y_train shape : {y_train_80.shape}")
print(f"y_test shape  : {y_test_20.shape}")

# Cek distribusi label untuk memastikan 'stratify' berjalan baik
print("\nDistribusi kelas sentimen (80:20):")
print("Data Train:\n", y_train_80.value_counts().sort_index())
print("Data Test :\n", y_test_20.value_counts().sort_index())
```

```
📊 SKENARIO 1: RASIO 80:20
X_train shape : (3621, 6447)
X_test shape  : (906, 6447)
y_train shape : (3621,)
y_test shape  : (906,)

Distribusi kelas sentimen (80:20):
Data Train:
 label_fix
Negatif     345
Netral      540
Positif    2736
Name: count, dtype: int64
Data Test :
 label_fix
Negatif     87
Netral     135
Positif    684
Name: count, dtype: int64
```

TRAIN TEST SPLIT (70:30)

```
# ================== VERSI 2: RASIO 70:30 ==================
X_train_70, X_test_30, y_train_70, y_test_30 = train_test_split(
    x_tfidf,
    y,
    test_size=0.3,            # 70% train, 30% test
    stratify=y,
    random_state=42
)

print("📊 SKENARIO 2: RASIO 70:30")
print(f"X_train shape : {X_train_70.shape}")
print(f"X_test shape  : {X_test_30.shape}")
print(f"y_train shape : {y_train_70.shape}")
print(f"y_test shape  : {y_test_30.shape}")

print("\nDistribusi kelas sentimen (70:30):")
print("Data Train:\n", y_train_70.value_counts().sort_index())
print("Data Test :\n", y_test_30.value_counts().sort_index())
```

```
📊 SKENARIO 2: RASIO 70:30
X_train shape : (3168, 6447)
X_test shape  : (1359, 6447)
y_train shape : (3168,)
y_test shape  : (1359,)

Distribusi kelas sentimen (70:30):
Data Train:
 label_fix
Negatif     302
Netral      473
Positif    2393
Name: count, dtype: int64
Data Test :
 label_fix
Negatif     130
Netral      202
Positif    1027
Name: count, dtype: int64
```

TRAIN MODEL RANDOM FOREST

```
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import (
    classification_report,
    confusion_matrix,
    ConfusionMatrixDisplay,
    accuracy_score,
    precision_score,
    recall_score,
    f1_score
)
import matplotlib.pyplot as plt
import pandas as pd

# ================== HYPERPARAMETER TERBAIK ==================
best_params = {
    'n_estimators': 800,
    'max_features': 'log2',
    'max_depth': None,
    'min_samples_split': 5,
    'min_samples_leaf': 1,
    'class_weight': 'balanced_subsample',
    'criterion': 'entropy',
    'ccp_alpha': 0.001,
    'bootstrap': True,
    'oob_score': True,
    'random_state': 42,
    'n_jobs': -1
}

print("✅ Section 1 Selesai: Library dan Hyperparameter siap digunakan!")
print("Hyperparameter yang digunakan:")
for k, v in best_params.items():
    print(f"   {k}: {v}")
```

```
✅ Section 1 Selesai: Library dan Hyperparameter siap digunakan!
Hyperparameter yang digunakan:
```

EVALUATION MODEL RANDOM FOREST

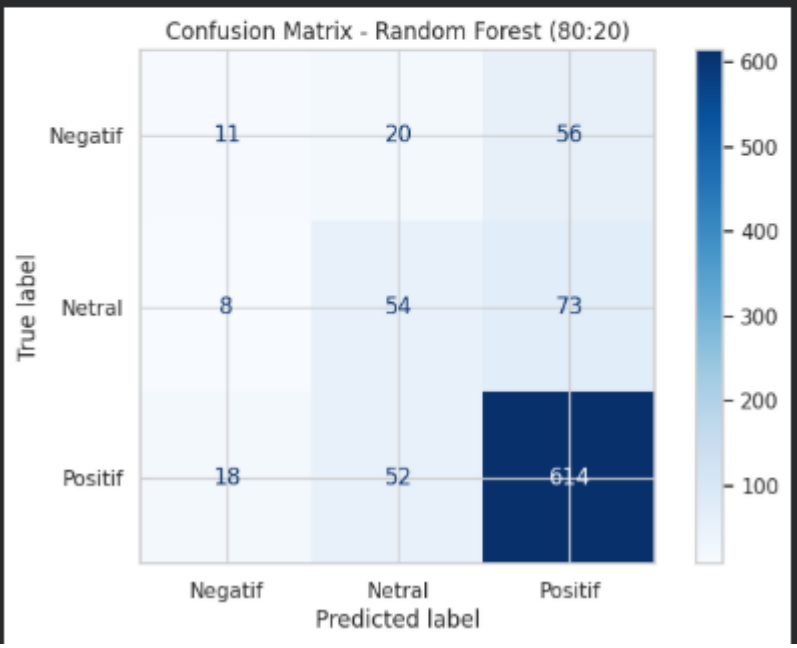
```
print("🚀 Memulai proses Training Random Forest (Tunggu beberapa saat)...")

# ================== TRAINING MODEL 80:20 ==================
print("\n⏳ Sedang melatih model rasio 80:20...")
rf_80 = RandomForestClassifier(**best_params)
rf_80.fit(X_train_80, y_train_80)
print("✔ Model 80:20 selesai dilatih!")

# ================== TRAINING MODEL 70:30 ==================
print("\n⏳ Sedang melatih model rasio 70:30...")
rf_70 = RandomForestClassifier(**best_params)
rf_70.fit(X_train_70, y_train_70)
print("✔ Model 70:30 selesai dilatih!")

print("\n" + "="*80)
print("✅ SEMUA PROSES TRAINING SELESAI!")
print("="*80)
```

```
🚀 Memulai proses Training Random Forest (Tunggu beberapa saat)...

⏳ Sedang melatih model rasio 80:20...
✔ Model 80:20 selesai dilatih!

⏳ Sedang melatih model rasio 70:30...
✔ Model 70:30 selesai dilatih!

================================================================================
✅ SEMUA PROSES TRAINING SELESAI!
================================================================================
```

CLASSIFICATION REPORT RASIO 80:20

```
print("="*80)
print("EVALUASI RANDOM FOREST - RASIO 80:20")
print("="*80)

# Prediksi menggunakan data test
y_pred_80 = rf_80.predict(X_test_20)

# Cetak Classification Report
print("Classification Report (80:20):")
print(classification_report(y_test_20, y_pred_80, digits=4, zero_division=0))

# Visualisasi Confusion Matrix
ConfusionMatrixDisplay.from_estimator(rf_80, X_test_20, y_test_20, cmap='Blues')
plt.title('Confusion Matrix - Random Forest (80:20)')
plt.show()

# Simpan metrik ke dalam dictionary untuk direkap nanti
hasil_80 = {
    'Rasio': '80:20',
    'Accuracy': accuracy_score(y_test_20, y_pred_80),
    'Precision_macro': precision_score(y_test_20, y_pred_80, average='macro', zero_division=0),
    'Recall_macro': recall_score(y_test_20, y_pred_80, average='macro', zero_division=0),
    'F1_macro': f1_score(y_test_20, y_pred_80, average='macro', zero_division=0),
    'OOB_Score': rf_80.oob_score_
}
```

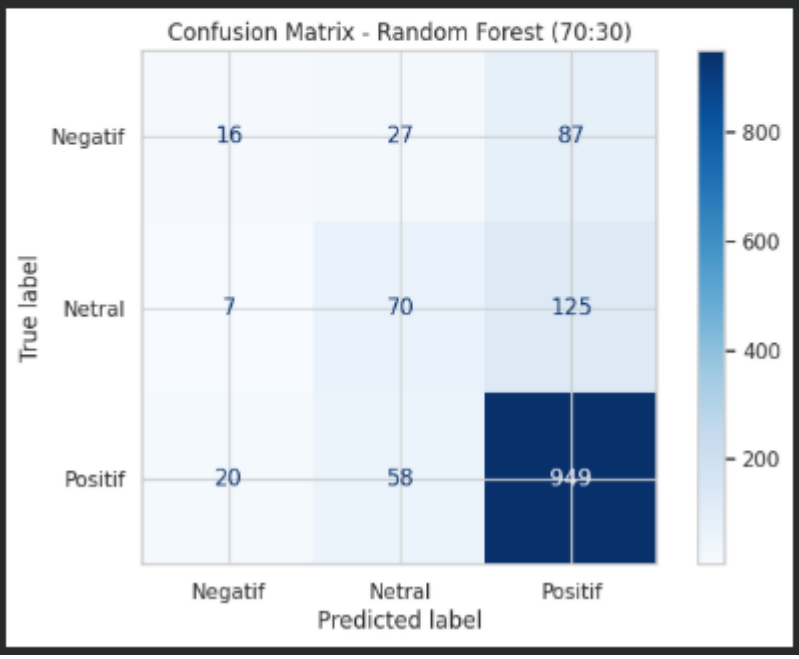
```
==========================================================================================
EVALUASI RANDOM FOREST - RASIO 80:20
==========================================================================================
Classification Report (80:20):
              precision    recall  f1-score   support

     Negatif     0.2973    0.1264    0.1774        87
      Netral     0.4286    0.4000    0.4138       135
     Positif     0.8264    0.8977    0.8605       684

    accuracy                         0.7494       906
   macro avg     0.5174    0.4747    0.4839       906
weighted avg     0.7163    0.7494    0.7284       906
```



CLASSIFICATION REPORT RASIO 70:30

```
print("="*80)
print("EVALUASI RANDOM FOREST - RASIO 70:30")
print("="*80)

# Prediksi menggunakan data test
y_pred_30 = rf_70.predict(X_test_30)

# Cetak Classification Report
print("Classification Report (70:30):")
print(classification_report(y_test_30, y_pred_30, digits=4, zero_division=0))

# Visualisasi Confusion Matrix
ConfusionMatrixDisplay.from_estimator(rf_70, X_test_30, y_test_30, cmap='Blues')
plt.title('Confusion Matrix - Random Forest (70:30)')
plt.show()

# Simpan metrik ke dalam dictionary
hasil_70 = {
    'Rasio': '70:30',
    'Accuracy': accuracy_score(y_test_30, y_pred_30),
    'Precision_macro': precision_score(y_test_30, y_pred_30, average='macro', zero_division=0),
    'Recall_macro': recall_score(y_test_30, y_pred_30, average='macro', zero_division=0),
    'F1_macro': f1_score(y_test_30, y_pred_30, average='macro', zero_division=0),
    'OOB_Score': rf_70.oob_score_
}

# ================== RINGKASAN PERBANDINGAN ==================
print("\n" + "="*90)
print("RINGKASAN PERFORMA RANDOM FOREST")
print("="*90)

# Gabungkan hasil 80:20 dan 70:30 menjadi satu DataFrame
summary_df = pd.DataFrame([hasil_80, hasil_70])
display(summary_df.round(4))

print(f"\n✅ Evaluasi selesai! Model siap dianalisis untuk bab hasil dan pembahasan skripsi.")
```

```
================================================================================
EVALUASI RANDOM FOREST - RASIO 70:30
================================================================================
Classification Report (70:30):
              precision    recall  f1-score   support

     Negatif     0.3721    0.1231    0.1850       130
      Netral     0.4516    0.3465    0.3922       202
     Positif     0.8174    0.9241    0.8675      1027

    accuracy                         0.7616      1359
   macro avg     0.5470    0.4646    0.4815      1359
weighted avg     0.7204    0.7616    0.7315      1359
```



WORDCLOUD SEMUA KOMENTAR

```
from wordcloud import WordCloud
import matplotlib.pyplot as plt

print("="*80)
print("WORDCLOUD - SELURUH KOMENTAR")
print("="*80)

# Menggabungkan seluruh teks dari kolom 'final_clean_text'
all_text = " ".join(df1['final_clean_text'])

# Membuat objek WordCloud
wordcloud_all = WordCloud(
    width=1200, height=700,
    background_color='white',
    colormap='viridis',
    max_words=350,
    min_font_size=10,
    max_font_size=160,
    contour_width=2,
    contour_color='black',
    random_state=42
).generate(all_text)

# Menampilkan hasil
plt.figure(figsize=(15, 8))
plt.imshow(wordcloud_all, interpolation='bilinear')
plt.axis('off')
plt.title('WordCloud - Seluruh Komentar Malaka Project', fontsize=18)
plt.show()
```


WORDCLOUD - SELURUH KOMENTAR
WordCloud - Seluruh Komentar Malaka Project
korupsi
pertamina
orang
hukum
bang
negara
indonesia
perintah
rakyat
jabat
bahan bakar
koruptor
usaha milik
hukum mati
kalo
badan usaha
milik negara
masyarakat
presiden
rugi
bakar minyak
salah
negeri
impor
harga
kaya
uang
hasil
turun
jaksa
oplos
triliun
untung
laku
kuasa
jokowi
jahat
pertamax
pertalite
penjara
biar
percaya
main
duit
hidup
coba
libat
tom
jual
dah
klo
tau
adil
beli
bener
jalan
gila
gitu
emang
sampe
bahas
hakim
moga
bilang
ron
langsung
bikin
keluarga
kerja
gua
politik
swasta
berantas korupsi
rugi negara
mati koruptor
bahasa bayi
ampas aset
prabowo
ahok
lembong
pakai
bongkar
blending
subsidi
paham
bangsa

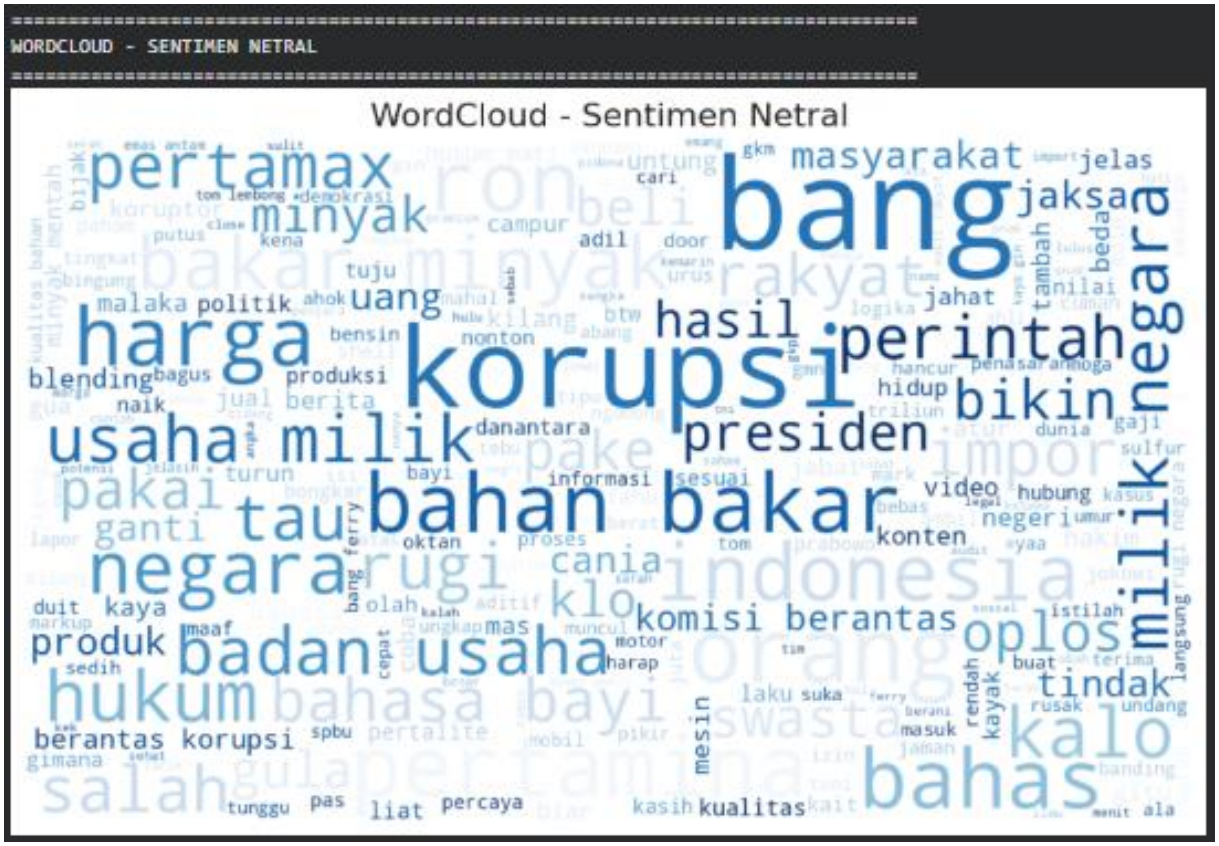
WORDCLOUD LABEL POSITIF

```
print("="*80)
print("WORDCLOUD - SENTIMEN POSITIF")
print("="*80)

# Memfilter data hanya untuk label 'Positif'
text_positif = " ".join(df1[df1['label_fix'] == 'Positif']['final_clean_text'])

# Membuat WordCloud Positif
wc_positif = WordCloud(
    width=1200, height=700,
    background_color='white',
    colormap='Greens',      # Tema warna hijau untuk sentimen positif
    max_words=250,
    min_font_size=10,
    max_font_size=130,
    random_state=42
).generate(text_positif)

# Menampilkan hasil
plt.figure(figsize=(12, 6))
plt.imshow(wc_positif, interpolation='bilinear')
plt.axis('off')
plt.title('WordCloud - Sentimen Positif', fontsize=16)
plt.show()
```


WORDCLOUD - SENTIMEN POSITIF
WordCloud - Sentimen Positif
korupsi
bahan bakar
negara
milik negara
hukum mati
rakyat
koruptor
bang
indonesia
usaha milik
perintah
pertamina
jabat
hukum
badan usaha
kerja
duit
gila
impor
langsung
rugi
jahat
presiden
jaksa
salah
bakar minyak
masyarakat
untung
trilliun
pertalite
tau
bilang
jokowi
moga
uang
kaya
tangkap
maling
emang
oplos
jalan
sampe
klo
tikus
doang
prabowo
swasta
cuman
berani
miskin
bikin
percaya
sakit
gimana
buka
bayar
anak
hilang
masuk
pilih
jujur
turun
tahan
busuk
subsidi
bijak
butuh
periksa
kualitas
bahas
gua
publik
putus
bangsa
harga
bahan
hasil
atas
pikir
sangka
dunia
tuntut
paham
malu
juta
mahal
gitu
suka
kuat
pas
adil
berantas korupsi
import
mudah
gaji
harap
baik
rezim
tindak
bagus
karna
tom

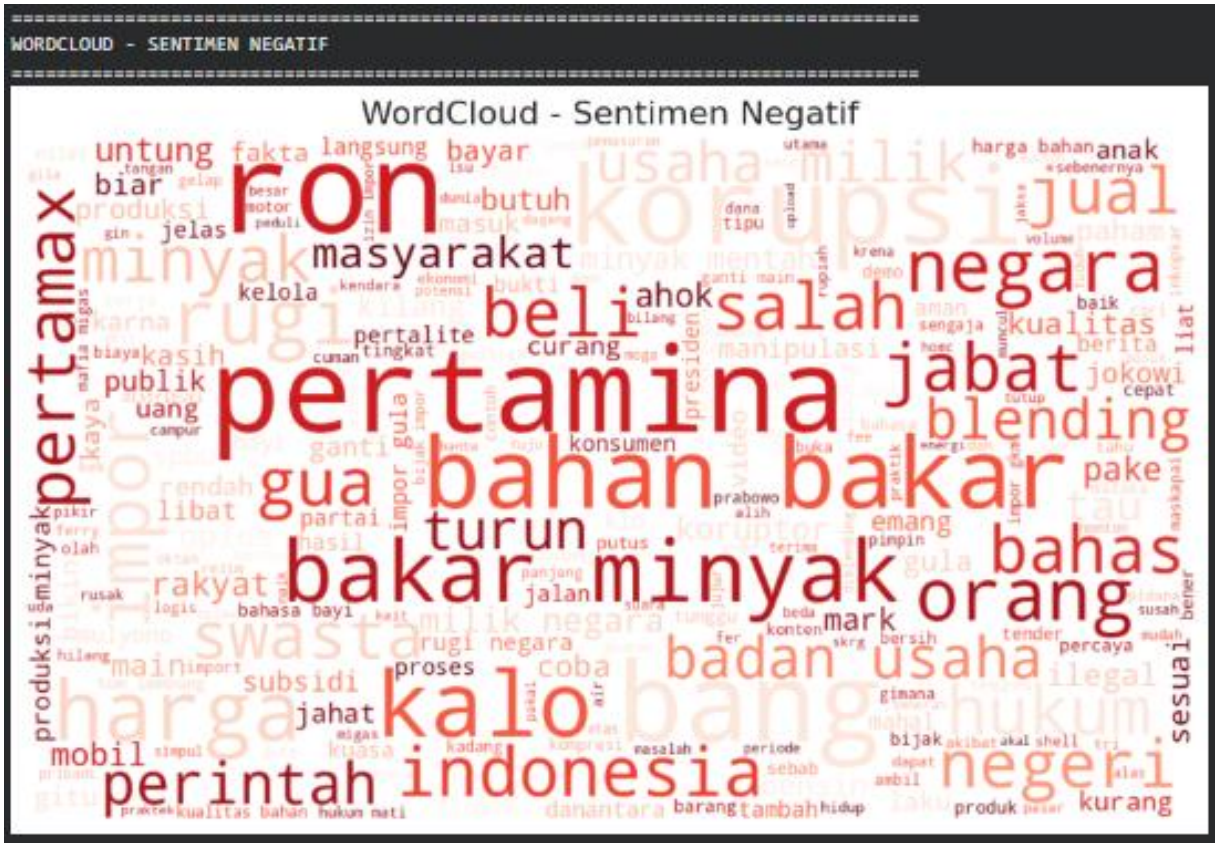
WORDCLOUD LABEL NETRAL

```
print("="*80)
print("WORDCLOUD - SENTIMEN NETRAL")
print("="*80)

# Memfilter data hanya untuk label 'Netral'
text_netral = " ".join(df1[df1['label_fix'] == 'Netral']['final_clean_text'])

# Membuat WordCloud Netral
wc_netral = WordCloud(
    width=1200, height=700,
    background_color='white',
    colormap='Blues',       # Tema warna biru untuk sentimen netral
    max_words=250,
    min_font_size=10,
    max_font_size=130,
    random_state=42
).generate(text_netral)

# Menampilkan hasil
plt.figure(figsize=(12, 6))
plt.imshow(wc_netral, interpolation='bilinear')
plt.axis('off')
plt.title('WordCloud - Sentimen Netral', fontsize=16)
plt.show()
```


WORDCLOUD - SENTIMEN NETRAL
WordCloud - Sentimen Netral
pertamax
minyak
bang
masyarakat
jaksa
negara
harga
korupsi
perintah
hasil
uang
bikin
usaha
milik
presiden
bahan bakar
tau
pakai
negara
rugi
indonesia
klo
komisi berantas
oplos
produk
badan usaha
hukum
bahasa bayi
tindak
kalo
berantas korupsi
salah
bahas
rakyat
jelas
jahat
blending
bagus
produksi
naik
danantara
informasi
sesuai
video
hubung
konten
ganti
oktan
duit
kaya
maaf
mas
motor
harap
sedih
buat
laku
suka
mesin
masuk
rusak
kayak
rendah
undang
tunggu
pas
liat
percaya
kasih
kualitas
gkm
cari
adil
door
tambah
beda
nilai
malaka
politik
ahok
tuju
bensin
nonton
btw
hidup
hancur
penasaran
moga
dunia
gaji
turun
bayi
proses
tom
negeri
umur
yaa
istilah
langsung
olah
cepat
ferry
spbu
pertalite
mobil
pikir
gimana
sehat
jangan
banding
menit
ala

WORDCLOUD LABEL NEGATIF

```
print("="*80)
print("WORDCLOUD - SENTIMEN NEGATIF")
print("="*80)

# Memfilter data hanya untuk label 'Negatif'
text_negatif = " ".join(df1[df1['label_fix'] == 'Negatif']['final_clean_text'])

# Membuat WordCloud Negatif
wc_negatif = WordCloud(
    width=1200, height=700,
    background_color='white',
    colormap='Reds',        # Tema warna merah untuk sentimen negatif
    max_words=250,
    min_font_size=10,
    max_font_size=130,
    random_state=42
).generate(text_negatif)

# Menampilkan hasil
plt.figure(figsize=(12, 6))
plt.imshow(wc_negatif, interpolation='bilinear')
plt.axis('off')
plt.title('WordCloud - Sentimen Negatif', fontsize=16)
plt.show()
```



CONTOH DATASET

teks	nomor_video	label_fix
Jangan lupa like, komen, dan share video ini supaya lebih banyak orang tahu!	video 1	Positif
KORUPTOR BANGSAT	video 1	Positif
SETIAP NEGARA PASTI PUNYA OLIGARKINYA MASING MASING,BEDANYA OLIGARKI NEGARA MAJU ADALAH INDUSTRIALISASI(mesin elektro tekno terbarukan chip dll)YG MENYANGGA EKONOMI NEGARA MAJU DI RANAH GLOBAL,DITAMBAH OLIGARKI DI NEGARA MAJU BERSETATUS CHEBOL(masih bisa di kendalikan pemerintah).SEMENTARA OLIGARKI DI NEGARA BERKEMBANG(indo)NON INDUSTRIALISASI(kaya dari tambang & impor)YANG MEMBUNUH KREATIFITAS INOVATIFITAS PRODUKTIFITAS RAKYAT ITU SENDIRI(oligarkinonindustri=racun negara).PARAHNYA OLIGARKI INDO SUDAH	video 1	Positif

teks	nomor_video	label_fix
TERLANJUR JADI MONSTER NAGA RAKSASA YG SULIT DI KENDALIKAN NEGARA(terbukti era habibi mega sby joko hanya bisa menyegel dalam artian hanya bisa diajak kerja sama tak bisa di perintah)BAHKAN SANGKING KAYANYA BISA MENGUASAI NEGARA ITU SENDIRI.KITA TAK BISA MENYALAHKAN PARA OLIGARKI BEGITU SAJA KARNA MEREKA BERJALAN SESUAI SISTEM HUKUM EKONOMI,YG SALAH ADALAH KEBIJAKAN YG DI AMBIL PEMERINTAHAN TERDAHULU.DIMANA SEMUA BERMULA DARI ORDE BARU YG SELAMA INI KITA ANGGAP PAHLAWAN TERNYATA MEMBERI SETENGAH NYAWA SAKTI NEGARA KPD SEGELINTIR CACING KECIL(keroni yg kelak jadi naga)DI TAHUN TAHUN EMAS(90an era negara negara hancur asia timur,Singapore,malaysia sedang pemulihan saling bersaing mulai industrialisasi dari nol,sementara soeharto terlena dngn pemasukan sda tambang migas membiarkan industri lokal mati kalah saing,uang negara hanya di prioritaskan ke para kroninya saja yg bergantung impor tanpa kreativitas produktifitas,rakyat di biarkan bodoh terlena dengan produk impor,tidak ada rasa ingin buat mobil lokal,semua hanya di doktrin jadi petani serentak tanpa pemasukan tambahkan dari teknologi)TAK TERASA 32 TAHUN BERKUASA KEMUDIAN CACING ITU DAH JADI NAGA RAKSASA DI PERPARAH 98 HUTANG NEGARA KIAN BERTAMBAH.GEN Z SEBAGAI GENERASI TERAKHIR DI SURUH NANGGUNG DOSA PARA LELUHUR(generasi kakek & bomer)UNTUK PERANG EKONOMI DENGAN TIRANI OLIGARKI DALAM NEGRI & TEKANAN IMPOR PERUSAHAAN LUAR NEGRI DENGAN MODAL SISA SISA KERAPUHAN SISTEM PEMERINTAHAN NEGRI INI YG DI GROGOTI PARA TIKUS TIRANI,ORMAS,PARTAI,HUTANG MENGGUNUNG,GODAAN KONTEN SAMPAH YG MERUSAK MORAL GEN Z ITU SENDIRI.TANPA BIMBINGAN,BEKAL ILMU YG MUMPUNI GEN Z DI PAKSA MENYEMBUHKAN NEGRI YANG SAKIT PARAH.MUNGKIN ITU SEBAB DI JULIKI GENERASI EMAS KARNA GENERASI AKHIR DEMOGRAFI		

teks	nomor_video	label_fix
INDONESIA,HARAPAN TERAKHIR PENYELAMAT EKONOMI NEGRI UNTUK GENERASI KEDEPAN HINGGA ANAK CUCU.JIKA ITU GAGAL NEGRI HANCUR,PORAK PORANDA PERANG SAUDARA,SEMENTARA BISNIS KONGLOMRAT MASIH BERDIRI TEGAK JADI PERUSAHAAN INTERNASIONAL.MAKANYA GEN Z INI JADI GENERASI PALING BERAT,PALING KERAS,PALING RAWAN YG AKAN MENGHADAPI BADAI EKONOMI YG DAHSYAT LEBIH DARI SEKEDAR KRISIS MONOTER 98.KALO TAK DI ARAHAN NEGRI INI AKAN SEPERTI AFRIKA DAN TIMTENG HANYA PERANG DAN PERANG UNTUK KEPENTINGAN.ycfc5f		
JIKA KALIAN INGIN INDONESIA MAJU SILAHKAN BACA INI. Selama ini kepala daerah bahkan presiden di jegal di kendalikan parpol yg terima duit tambahan kampanye dll dari para oligarki.partai tak salah tapi sistem yg salah. SISTEM PEMERINTAHAN SALAH=ORANG-ORANGNYA PUN IKUT SALAH. A> SISTEM INDO HARUS DI UBAH JADI:Demokrasi Terpimpin Modern(bukan Orde Lama),(Ini BUKAN otoriter,BUKAN militeristik,tapi:Demokrasi dibatasi,Kekuasaan dipusatkan,Parpol dikerdilkan).artinya: *Presiden sangat kuat(tapi diawasi).Presiden dipilih langsung rakyat.Masa jabatan1x panjang(8â€“10 tahun),>Hasilnya supaya fokus kerja bukan kampanye,Tidak bisa dijatuhkan DPR kecuali:Korupsi berat,Pengkhianatan negara. *DPR tidak dominan.DPR tidak bisa sandera kebijakan,Tidak ada bagi-bagi kursi menteri,Fungsi DPR dipersempit(Pengawasan Anggaran,Persetujuan terbatas UU strategis).>hasilnya DPR bukan pengendali negara,hanya pengawas. B> TRANSISI 10 TAHUN:â€œMENETRALKAN PARPOLâ€• .(Ini bagian paling krusial). *Tahun1â€“3:Penjinakan.Audit keuangan seluruh parpol,Dana swasta besar dilarang,Parpol hanya boleh terima dana negara.Pelanggaran=dibekukan.> hasilnya 50% parpol akan gugur alami. *Tahun4â€“6:Kompetisi	video 1	Positif

teks	nomor_video	label_fix
independen.Calon independen dibuka luas.DPR diisi campuran(Independen & Parpol lolos audit).>hasilnya Parpol bukan satu-satunya pintu kekuasaan. *Tahun7â€“10:Parpol sebagai klub ide.Parpol tidak wajib ikut pemilu,Tidak punya hakâ€œjatah jabatanâ€• ,Hanya berfungsi:Pendidikan politik,Rekrut kader.>hasilnya Kalau rakyat tak percaya:mati sendiri. C> BIROKRASI,TEKNOKRAT=JANTUNG NEGARA.Negara tidak boleh tergantung politisi. *Menteri & pejabat=profesional(Dipilih lewat:Rekam jejak,Uji publik,Panel independen,Parpol dilarang titip jabatan teknis). *Lembaga kunci super kuat(KPK,MA,BPKâ†'otonom penuh,Gaji tinggi,Hukuman) pelanggaran=sosial+hukum.>hasilnya Elite takut hukum,bukan sebaliknya. D> DEMOKRASI LANGSUNG TERBATAS(aman).Bukan semua rakyat voting tiap hari(itu chaos).Yg boleh referendum:Amandemen UUD,Penjualan SDA strategis,Utang besar negara,Pemindahan ibu kota.>hasilnya isu besar:rakyat langsung. E> ELITE PASTI TAKUT MODEL INI KARNA:Hilang ladang uang & posisi aman,Tidak bisa sandera presiden & barter proyek.Ingat Parpol hari ini:Bukan alat rakyat,tapi alat bertahan hidup elite. F> PARPOL DIBUBARKAN TOTAL=BERESIKO.Jika parpol dihapus tanpa desain:Militer naik,Intel dominan,Kritik disamakan makar,Media dikendalikan.> Sebab itu parpol jangan dibunuh,tapi dikurung. G> INGAT.*Parpol bukan sumber demokrasi.Rakyat+hukum+birokrasi bersih=Demokrasi sejati.Masalah Indo sekarang:Demokrasi prosedural bukan Demokrasi substansi.-Solusi:Kurangi aktor politik,perkuat negara,lindungi rakyat.-Terakhir (ini penting),Banyak orang merasa:Apatis,Marah,Ingin sistem diganti total.Itu bukan tanda bodoh,tapi tanda sistem gagal memberi harapan.txdd5x		
12:33	video 1	Negatif

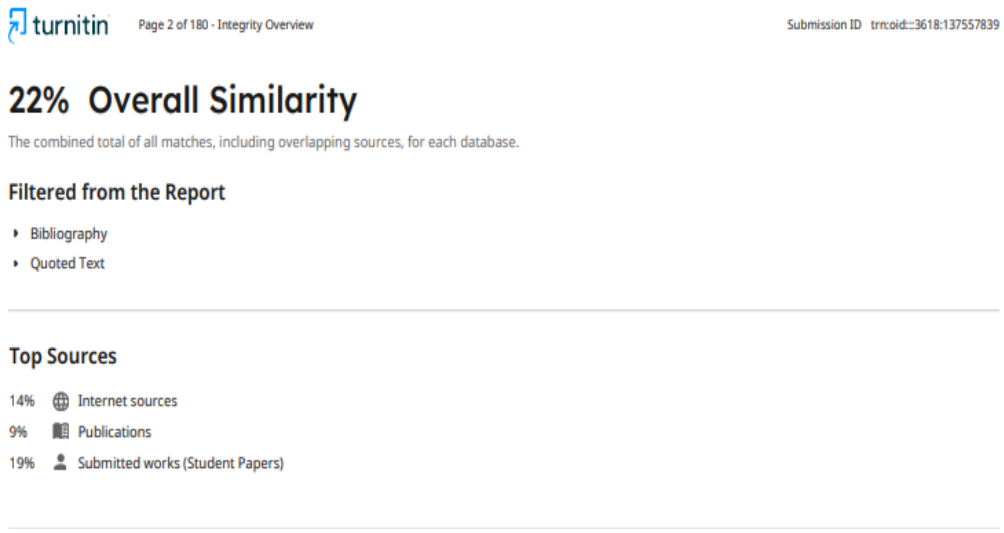
teks	nomor_video	label_fix
KAWAL.PENYELIDIKAN AGAR TDK MASUK ANGIN. KARNA YG MEREKA.LAKUKAN BENER2.MENYENGSARAKAN.RAKYAT DGN MERUSAK EKONONOMI.	video 1	Positif
Ini sudah berlangsung puluhan tahun ... pertamina itu Bobrok.., hanya ucapan ini yg bisa saya sampaikan.. aneh ajja kok Pak Prabowo nggak berani mencopot si Bahlil ya.. ??? Ada apa ini...	video 1	Positif
Gila banget orang2 Pertamina yg kemaruk duit ini!	video 1	Positif
Sehat sukses slalu nak Ferry Irawan skluarga besr dn mahasiswa buruh rkyt lndonesia kedepan ALLAH SWT TDK tdur setiap baik bruk prbuatan kita sendri yg trima	video 1	Positif
Indonesia kaya bnyk orng crds di bidang nya hmmm jdii apa yang g terjadi ya rkyt hnya diam KRNA mng TDK mngerti tau takut dn mngerti TPI TDK bisa brbuat bnyk hnya doa yg trbaik tuk anak cucu gnersi kdepn lndonesia lebih baik TDK rakussz disaat rkyt lapar bila TDK bekerja mncari cuan saat sulitnya ekonomi zmn ni	video 1	Positif
Harusnya rakyat bisa mengajukan Classaction ke Negara , karena sebagai konsumen pertamina rakyat dirugikan	video 1	Positif
Bajingan ya kita ditipu bertahun tahun duit segitu gedenya	video 1	Positif
Coba feri tanya ahok kenapa dia membiarkan korupsi di pertamina,sedangkan dia punya bukti bukti malah di endapkan ,untungnya jaksa berani.	video 1	Positif
Yg paling jahat lagi korupsi haji bang,karna itu sudah masuk dalam agama Islam,tolong dong d bahas.	video 1	Positif
Mantap bung ferry arwandi bongkar kasus korupsi borok para pejabat maling2 berdasi tdk empati dgn rakyanya sendiri	video 1	Positif
Orang waras bernasehat : Cari orang pintar itu sekarang gampang, tapi cari yang jujur susah. Orang waras tapi di Indonesia : Jadi orang jujur emang bagus dan disayang Tuhan, tapi musuhnya kegedhean. ðŸ˜Š	video 2	Positif
Awesomely .. thanks mbak .. sehat selalu	video 2	Positif
Laaaa muncul juga ni orang sok penting ini ðŸ¥±ðŸ¥±ðŸ¥±	video 2	Netral
Indonesia dgn populasi 285 juta jiwa tapi terlalu kebanyakkan SDM rendahnya, sehingga kalap lama2	video 2	Netral

teks	nomor_video	label_fix
Kurang sedekah, kurang solat, kurang ngaji, kurang literatur bahasa timur tengah, makanya kena kasus????	video 2	Netral
Saya kira video ini mau bahas vonis TL secara detail, tapi bikin infografis timeline atau case position saja tidak, kecewa saya. Mbak Cania atau tim produksinya sudah baca vonisnya? atau tanya sama yang paham hukum? di Malaka kan ada yg lulusan FHUI, juga kemarin anda baru saja ngobrol sama Dr. Pram, owner law firm terkenal lulusan FHUI juga yg pinter dan sudah ambis before it's trendy, mungkin dia bisa kasih insight dari perspektif hukum. Kenapa penting untuk baca vonisnya dulu? karena pertanyaan-pertanyaan di video ini sudah dijawab oleh fakta persidangan dan pertimbangan hakim dalam vonis tersebut. Saya sedang baca vonisnya, belum selesai masih di halaman 1300-an dari 1561 halaman, tapi garis besarnya sudah dapet lah. Misalnya anda menanyakan kenapa impor GKM disalahkan? karena hasil rakortas (2016 kalo gak salah) adalah utk mengimpor GKP, bukan GKM, jadi kebijakan impor GKM itu tidak tepat dan patut dipertanyakan motifnya karena kalau tujuannya mengatasi kelangkaan gula pasir di masyarakat sedangkan yg diimpor adalah GKM yg masih butuh waktu utk pengolahan beberapa bulan sampai jadi GKP maka problem kelangkaan jadi tidak teratasi secara tepat waktu. Ada juga pertanyaan soal kenapa TL berulang kali memperpanjang izin Inkopkar utk melakukan operasi pasar padahal performancenya terbukti jelek. Inkopkar gagal memenuhi target operasi pasar yg seharusnya hanya berlangsung dalam rangka Lebaran 2015, tapi hanya mampu mendistribusikan 50% dari target, lalu minta perpanjangan ke mendag Rachmat Gobel tapi sampai akhir jabatannya tidak diberikan (mungkin karena performancenya jelek itu). Baru setelah Rachmat Gobel turun dan diganti TL lalu TL menerbitkan SK perpanjangan utk Inkopkar hingga akhir tahun 2015 dan diperpanjang lagi hingga pertengahan 2016. Apakah wajar rekanan yg performanya jelek bukannya dievaluasi utk dihentikan kerjasamanya malah diperpanjang berkali-kali? Kenapa kerjasama dgn Inkopkar ini jadi bermasalah? karena gula yg didistribusikan Inkopkar bukan milik Inkopkar sendiri, melainkan dia ambil dari PT AP, lalu Inkopkar minta kompensasi ke kemendag agar menerbitkan izin impor	video 2	Negatif

teks	nomor_video	label_fix
GKM utk PT AP, dgn maksud mengganti gula dari PT AP tadi (100rb ton diganti 105rb ton, ada kelebihan 5000 ton tuh). Makanya kebijakannya jadi impor GKM padahal ketika Rakortas akhir 2015 atau awal 2016 membolehkan impor GKP yg diimpor justru GKM. Selanjutnya yg ikutan jadi Importir GKM ini selain PT AP ada sekitar 7-8 perusahaan swasta lainnya. Mereka inilah pihak-pihak yg mendapatkan keuntungan dari kebijakan impor GKM (yang tidak tepat, karena pada Rakortas Mei 2015 dipaparkan data yg menyatakan stok gula cukup jadi tidak usah impor) yg dikeluarkan TL, sehingga unsur Pasal 2 UU Tipikor soal "memperkaya diri sendiri atau orang lain" terpenuhi. Keuntungan dari mana? karena impor di harga 8900 rupiah/kg lalu dijual lagi 9105 rupiah/kg (5 rupiahnya bagian utk Sucofindo), jadi 8 perusahaan swasta ini ada untung 200 rupiah/kg dari jatah impor ratusan ribu Ton itu. Makanya kemudian Laporan Audit BPKP menemukan adanya kerugian 500 milyar lebih dari importasi gula ini. Anak hukum utk memahami kasus yg diberikan maka dia harus bikin case positionnya dulu, di-summarize biar cerita yang rumit bisa lebih sederhana. Dibikin kronologis/timelinenya dulu, baru bisa dikaji dan disimpulkan ada kasus atau tidak, kalau ada kasus ini kenanya pasal mana, dst. Bukan berangkat dari pertanyaan-pertanyaan anekdotal yg lompat di tengah cerita tanpa memahami awalnya dulu. Pertanyaan-pertanyaan di video ini terasa seperti bukan dari orang yg sudah berusaha membaca putusan kasusnya melainkan seperti mengulang argumen tim hukum TL, entahlah, apalagi baik channel Malaka maupun Ferry Irwandi sama-sama sedang membahas ini.		
jurus Ninja Jokowi apa yang dia katakan harus terjadi.	video 2	Positif
Preseden buruk	video 2	Positif
Dia dukun yang nyamar jadi jaksa	video 2	Positif
Ini bukan tentang mendukung siapa atau yang lain tapi ini tentang keadilan Tom lembong tidak bersalah jelas ini fitnah	video 2	Positif
BUMN kita terbiasa dengan keuntungan hayalan, misal flag carrier nasional kita, forecast profit 10T pencapaian 7T kita akan bilang rugi 3T bukan profit 7T...demikian konsep negara ini.	video 2	Positif

teks	nomor_video	label_fix
sepanjang persidangan, Tom Lembong dituding dan dihukum maladministratif (bodoh) oleh JPU & Hakim, tiba2 TL divonis korupsi (jahat), ga nyambung gw yakin banget 99%, Jaksa dan Hakim di persidangan TL ini korup, karena mens rea nya udh ada, 1% nya tinggal pembuktian aja	video 2	Positif
Kok kita jadi takut ngambil kebijakan, walaupun kebijakan itu untuk rakyatðŸ˜¢	video 2	Positif
Aneh bgt kk, trus liat kasus gini, kita harus nya gmn kk,	video 2	Netral
jadi menteri PROJudOl aja aman	video 2	Positif
Kalu gitu yuk tuntut Jokowi karena udah menjual IKn dgn HGU ratusan tahun karena sangat mendukung kapitalisme	video 2	Positif

LAMPIRAN SIMILARITY REPORT TURNITIN


turnitin Page 2 of 180 - Integrity Overview Submission ID trn:oid:::3618:137557839

22% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Filtered from the Report

- Bibliography
- Quoted Text

Top Sources

14% Internet sources
9% Publications
19% Submitted works (Student Papers)

LAMPIRAN TERJEMAHAN BAHASA INGGRIS

ABSTRACT

Alfathir, Muhammad Dimas. 2026. **The Sentiment Analysis on Youtube Viewer Comments on Corruption Topic on Malaka Project Educational Channel Using Random Forest Algorithm. Thesis. Library and Informatics Science, Faculty of Science and Technology Universitas Islam Negeri Maulana Malik Ibrahim Malang. Advisors: (I) Fakhris Khusnu Reza Mahfud, M. Kom. (II) Ahmad Yulianto, M.Pd.I.**
Keywords: Sentiment Analysis, Random Forest, TF-IDF, Corruption, Malaka Project.

YouTube social media has become a digital space to express opinions on public issues, including corruption. The research aims to analyze the sentiment classification of viewer comments on two educational videos published by the Malaka Project channel, namely "Korupsi Terstruktur dan Terjahat Pertamina" and "Vonis Tom Lembong Adalah Masalah Publik". It also compares the sentiment distribution across the two videos. It employed a quantitative approach based on text mining using the Random Forest algorithm. The researcher collected 4,544 comments through scraping techniques and involved three language experts to manually label the comments into three categories—positive, neutral, and negative. The research procedures consisted of text preprocessing, term weighting using TF-IDF, splitting the dataset into training and testing sets with ratios of 80:20 and 70:30, and evaluating the model using a confusion matrix. The research results indicate that the 70:30 ratio scenario generated the best performance, achieving 76% accuracy, 72% weighted average precision, 76% recall, and a 73% f1-score. However, the macro average values remained low, with 54% precision, 46% recall, and a 48% f1-score, due to the imbalanced label distribution dominated by positive sentiment. In the first video, positive sentiment reached 75% of the comments, while in the second video it reached 78.6%. The proportion of negative sentiment in the case of individual corruption was only 3%, lower than the 11% observed in systemic corruption cases. These findings suggest that viewers tend to show more defensive attitude toward cases of individual corruption. Therefore, the research recommends the application of imbalance-handling techniques, the expansion of data sources, and comparisons with other algorithms to improve the accuracy of minority sentiment detection while still considering the Indonesian language context and the digital conversation dynamics.

Penerjemah	Tanggal
Rizka Yaniary NIPPPK. 199801242023212005	13-05-2026

LAMPIRAN TERJEMAHAN BAHASA ARAB

مستخلص البحث

الفاطر، محمد دماس. ٢٠٢٦. تحليل المشاعر لتعليقات مشاهدي يوتيوب حول موضوع الفساد في قناة مشروع مالاكا التعليمي باستخدام خوارزمية الغابة العشوائية. البحث الجامعي. قسم المكتبات وعلوم المعلومات، كلية العلوم والتكنولوجيا بجامعة مولانا مالك إبراهيم الإسلامية الحكومية مالانج. المشرف الأول: فخريس لحسن رضا محفوظ، الماجستير؛ المشرف الثاني: أحمد يوليانتو، الماجستير.

الكلمات الرئيسية: تحليل مشاعر، غابة عشوائية، TF-IDF، فساد، مشروع مالاكا.

أصبح وسائل التواصل الاجتماعي يوتيوب مساحة رقمية للمجتمع للتعبير عن آرائهم بشأن القضايا العامة، بما في ذلك الفساد. يهدف هذا البحث إلى تحليل تصنيف المشاعر لتعليقات المشاهدين على مقطعين فيديو لقناة مشروع مالاكا التعليمي، وهما "الفساد المنظم والأجرم لشركة فرتامينا" و"حكم توم لمبونج هو قضية عامة"، وكذلك مقارنة توزيع المشاعر بين هذين المقطعين. استخدم هذا البحث منهجا كميا مبنيا على استخراج النصوص باستخدام خوارزمية الغابة العشوائية. تم الحصول على البيانات من خلال تقنية الاستخراج وعددها ٤٥٤٤ تعليقًا، ثم تم وضع العلامات يدويًا من قبل ثلاثة خبراء لغويين إلى ثلاث فئات، وهي الإيجابية، والمحايدة، والسلبية. تشمل مراحل الدراسة معالجة النصوص المسبقة، وتوزين الكلمات باستخدام تي اف-اي دي دف (TF-IDF)، وتقسيم البيانات إلى بيانات تدريبية وبيانات اختبار بنسبة ٨٠:٢٠ و٧٠:٣٠، وكذلك تقييم النموذج باستخدام مصفوفة الارتباك. أظهرت نتائج البحث أن سيناريو النسبة ٧٠:٣٠ يعطي أفضل أداء بدقة ٧٦٪، ومتوسط الوزن للثبات ٧٢٪، والاسترجاع ٧٦٪، ودرجة ف١ ٧٣٪. ومع ذلك، فإن قيمة المتوسط الكلي لا تزال منخفضة، حيث الثبات ٥٤٪، والاسترجاع ٤٦٪، ودرجة ف١ ٤٨٪، بسبب عدم توازن توزيع التصنيفات وسيطرة المشاعر الإيجابية. في الفيديو الأول، وصلت نسبة المشاعر الإيجابية إلى ٧٥٪، بينما في الفيديو الثاني بلغت ٧٨٫٦٪. ونسبة المشاعر السلبية في حالات الفساد الفردي هي فقط ٣٪، وهي أقل من الفساد النظامي الذي يبلغ ١١٪. تشير هذه النتائج إلى أن المشاهدين يميلون إلى اتخاذ موقف أكثر دفاعية في مواجهة حالات الفساد الفردية. يوصي هذا البحث باستخدام تقنيات معالجة البيانات غير المتوازنة، وتوسيع مصادر البيانات، وكذلك مقارنة الخوارزميات الأخرى من أجل جعل اكتشاف مشاعر الأقلية أكثر دقة مع مراعاة سياق اللغة الإندونيسية وديناميكيات المحادثات الرقمية.

Penerjemah	Tanggal
M.Mubaasyir Munir, MA NIPPPK. 198605132022211010	13-05-2026