SENAFTI
SEMINAR NASIONAL MAHASISWA
FAKULTAS TEKNOLOGI INFORMASI

6th Seminar Nasional Mahasiswa Fakultas Teknologi Informasi (SENAFTI)
20 September 2025 – Jakarta, Indonesia
Volume 4, Nomor 2, September 2025 - ISSN 2962-8628 (*online*)

ANALISIS SENTIMEN KOMENTAR YOUTUBE TERHADAP ISU BISNIS GELAP DOKTER DAN PERUSAHAAN FARMASI MENGGUNAKAN ALGORITMA *NAÏVE BAYES*

Septian Farriz Hartono[1*], Achmad Solichin[2], Noni Juliansari[3], Purwanto[4]

[1,2,3,4]Teknik Informatika, Fakultas Teknologi Informasi, Universitas Budi Luhur, Jakarta Selatan, Indonesia

Email: [1*]1911510624@student.budiluhur.ac.id, [2]achmad.solichin@budiluhur.ac.id, [3]noni.juliasari@budiluhur.ac.id, [4]purwanto@budiluhur.ac.id.
(* : corresponding author)

Abstrak-Perkembangan media sosial, khususnya YouTube, menjadikannya sebagai sarana utama masyarakat untuk menyuarakan opini terhadap berbagai isu sosial, termasuk praktik bisnis gelap yang melibatkan dokter dan perusahaan farmasi. Permasalahan yang muncul adalah bagaimana mengklasifikasikan sentimen masyarakat dari komentar yang beragam, tidak baku, dan seringkali ambigu. Penelitian ini bertujuan menganalisis sentimen komentar pada kanal YouTube MALAKA menggunakan algoritma *Naïve Bayes*, yang dikenal sederhana namun efektif dalam klasifikasi teks berbasis probabilitas. Data dikumpulkan melalui proses *crawling*, kemudian diproses dengan tahapan *case folding, tokenisasi, stopword removal, stemming,* dan ekstraksi fitur menggunakan *TF-IDF*. Pengolahan data dilakukan dengan bahasa pemrograman *python* dan pustaka pendukung seperti. Sentimen dikategorikan ke dalam dua kelas, yaitu positif dan negatif, dengan skema pembagian 80% data latih dan 20% data uji. Hasil pengujian menunjukkan bahwa algoritma *Naïve Bayes* mencapai akurasi sebesar 88%, dengan presisi dan recall tinggi, khususnya pada sentimen negatif (82% dan 100%) serta positif (95% dan 81%). Temuan ini mengindikasikan dominasi komentar bernuansa negatif yang mencerminkan ketidakpercayaan publik terhadap etika profesi medis dan integritas perusahaan farmasi.

Kata Kunci: analisis sentimen, youtube, *naïve bayes*.

SENTIMENT ANALYSIS OF YOUTUBE COMMENTS ON THE ISSUE OF ILLICIT BUSINESS PRACTICES BY DOCTORS AND PHARMACEUTICAL COMPANIES USING THE NAÏVE BAYES ALGORITHM

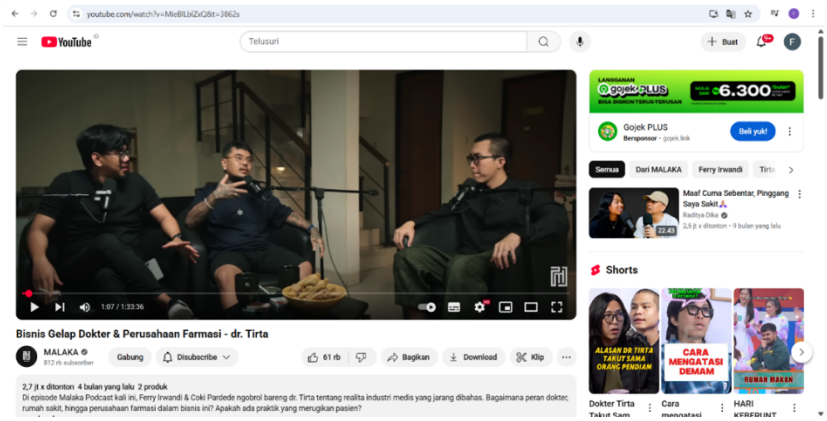
Abstract- *The development of social media, particularly YouTube, has made it a primary platform for the public to voice opinions on various social issues, including illicit business practices involving doctors and pharmaceutical companies. The main problem that arises is how to classify public sentiment from comments that are diverse, non-standard, and often ambiguous. This study aims to analyze sentiments from comments on the MALAKA YouTube channel using the Naïve Bayes algorithm, which is known for its simplicity yet effectiveness in probabilistic text classification. Data were collected through a crawling process and then processed through several stages, namely case folding, tokenization, stopword removal, stemming, and feature extraction using TF-IDF. Data processing was carried out using the Python programming language with supporting libraries. Sentiments were categorized into two classes, positive and negative, with an 80% training data and 20% testing data split. The results show that the Naïve Bayes algorithm achieved an accuracy of 88%, with high precision and recall, particularly for negative sentiment (82% and 100%) and positive sentiment (95% and 81%). These findings indicate the dominance of negative comments, reflecting public distrust toward medical ethics and the integrity of pharmaceutical companies.*

Keywords*: sentiment analysis, youtube, naïve bayes algorithm.*

1. PENDAHULUAN

Dalam era digital saat ini, media sosial dan *platform* daring telah menjadi wadah utama bagi masyarakat untuk mengekspresikan opini, pandangan, dan perasaan terhadap berbagai isu, produk, maupun layanan. Jumlah data yang dihasilkan setiap harinya sangat besar dan terus meningkat, sehingga mendorong perlunya suatu metode yang efektif untuk menggali informasi berharga dari data tersebut. Salah satu pendekatan yang banyak digunakan adalah analisis sentimen, yaitu proses mengidentifikasi dan mengkategorikan opini yang diekspresikan dalam suatu teks, terutama untuk menentukan apakah sikap penulis terhadap suatu topik bersifat positif, negatif, atau netral [1].

YouTube menjadi salah satu platform media sosial yang banyak digunakan oleh masyarakat untuk mengekspresikan opini secara terbuka melalui kolom komentar. Dalam konteks isu-isu sensitif, seperti dugaan adanya bisnis gelap antara dokter dan perusahaan farmasi, komentar dari pengguna YouTube mencerminkan beragam pandangan publik. Bisnis gelap ini merujuk pada praktik tidak etis di mana sebagian oknum dokter diduga bekerja sama dengan perusahaan farmasi untuk meresepkan obat tertentu demi keuntungan pribadi, bukan berdasarkan kebutuhan medis pasien. Praktik seperti ini menimbulkan kekhawatiran masyarakat karena dapat mengorbankan keselamatan pasien dan merusak integritas profesi kedokteran serta industri farmasi [2]. Banyak video yang mengangkat isu ini di YouTube memancing reaksi keras dari warganet, yang kemudian dituangkan melalui kolom komentar sebagai bentuk keresahan dan kritik terhadap sistem kesehatan yang dianggap korup [3].



Gambar 1. Halaman YouTube Malaka Project

Di episode Malaka Podcast kali ini, Ferry Irwandi dan Coki Pardede berbincang bersama dr. Tirta untuk membedah realita industri medis yang sering luput dari perhatian publik. Mereka mengulas secara kritis peran dokter, rumah sakit, hingga perusahaan farmasi dalam dinamika bisnis kesehatan, serta membahas kemungkinan adanya praktik-praktik yang merugikan pasien [4]. Keyakinan mereka bahwa akses terhadap pendidikan yang berkualitas dapat membentuk "Masyarakat Baru" masyarakat yang cerdas, kritis, dan empatik menjadi latar penting dari diskusi ini [5]. Hal ini terbukti dari antusiasme publik yang tinggi, dengan jumlah komentar yang mendekati 4.000 dan penayangan yang telah melampaui 2,5 juta kali, menunjukkan bahwa isu yang diangkat berhasil menggugah kesadaran dan perhatian masyarakat luas.

Namun, menganalisis sentimen dari komentar di YouTube bukanlah hal yang sederhana. Tantangan utama dalam proses ini adalah adanya penggunaan bahasa informal, singkatan, campuran bahasa, serta ekspresi emosional seperti sarkasme dan ironi [6]. Keragaman gaya bahasa dan cara penyampaian opini oleh pengguna sering kali menyulitkan sistem analisis otomatis untuk memahami maksud sebenarnya dari suatu komentar. Terlebih lagi, komentar yang membahas isu kontroversial seperti bisnis gelap antara dokter dan perusahaan farmasi sering kali bersifat emosional dan tajam, sehingga diperlukan pendekatan analisis yang mampu menangkap nuansa makna secara akurat [7].

Beberapa penelitian sebelumnya telah berupaya menganalisis berbagai data media sosial. Pada penelitian yang dilakukan oleh [8], dilakukan analisis sentimen masyarakat terhadap Ustadz Abdul Somad melalui komentar YouTube menggunakan algoritma *Naïve Bayes*. Dari 1000 komentar pada 10 video, diperoleh 67% sentimen positif, 27% netral, dan 6% negatif. Algoritma *Naïve Bayes* dipilih karena sederhana dan akurat, dengan hasil evaluasi menunjukkan akurasi yang baik. Hasil tersebut menunjukkan bahwa metode *Naive Bayes* efektif untuk analisis sentimen. Selanjutnya, penelitian yang dilakukan oleh [8], menganalisis sentimen masyarakat terhadap sistem e-Tilang menggunakan data komentar dari YouTube. Dengan mengumpulkan 500 komentar, data diproses melalui tahap *preprocessing* untuk mengurangi *noise*. Algoritma *Naïve Bayes* digunakan untuk mengklasifikasikan sentimen tanpa memerlukan pemodelan statistik. Hasil penelitian menunjukkan akurasi sebesar 79,44%, dengan penerapan seleksi fitur yang mengoptimalkan tingkat akurasi analisis sentimen terhadap sistem e-Tilang. Selanjutnya penelitian oleh [10], menganalisis sentimen masyarakat Indonesia terhadap mobil listrik melalui komentar di YouTube menggunakan algoritma *Naïve Bayes* dan pendekatan KDD. Rendahnya pemahaman publik menjadi tantangan utama yang memicu sentimen negatif. Dengan penerapan teknik *SMOTE Upsampling*, akurasi model meningkat dari 50,70% menjadi 70,69%, meskipun presisi dan *recall* masih perlu ditingkatkan. Hasil penelitian menunjukkan bahwa *Naïve Bayes* memiliki potensi dalam analisis sentimen, dan disarankan untuk memperluas data, menguji algoritma lain, serta melakukan pengujian lebih lanjut guna meningkatkan akurasi dan penerimaan mobil listrik di Indonesia.

Tabel 1. Studi Literatur

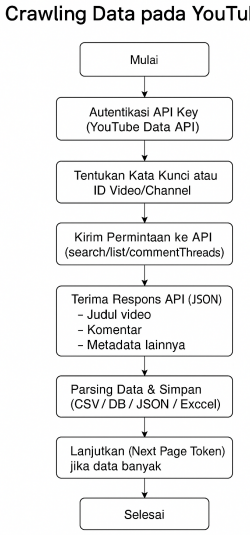
Tahun	Paper	Tujuan Penelitian	Hasil Penelitian
2020	[5]	Penelitian ini bertujuan untuk mengembangkan aplikasi Android berbasis YouTube API sebagai media pembelajaran digital bagi peternak burung kenari di Jawa Timur. Aplikasi ini dirancang untuk memudahkan akses informasi edukatif melalui video YouTube guna mendukung peningkatan kualitas usaha penangkaran kenari.	Aplikasi dikembangkan menggunakan metode *waterfall*, diuji dengan *black-box* dan beta *testing*, serta dinyatakan berjalan baik dan mudah digunakan. Dilengkapi tujuh fungsi utama seperti pencarian video, pemutaran, penyimpanan favorit, hingga integrasi YouTube API dengan Retrofit dan *database* Realm, aplikasi ini dinilai efektif dalam memberikan informasi praktis dan mendukung pengembangan usaha peternak kenari.
2022	[2]	Penelitian ini bertujuan untuk menganalisis sentimen masyarakat terhadap Ustadz Abdul Somad berdasarkan komentar pengguna pada platform YouTube dengan menggunakan algoritma *Naïve Bayes*. Metode ini dipilih karena dikenal sederhana namun memiliki tingkat akurasi yang baik dalam klasifikasi teks	Hasil penelitian menunjukkan bahwa dari 1.000 komentar yang dianalisis pada 10 video YouTube, terdapat 67% sentimen positif, 27% netral, dan 6% negatif. Algoritma *Naïve Bayes* yang digunakan menghasilkan tingkat akurasi sebesar 87%, *precision* 91%, *recall* 97%, dan *F-measure* 93%. Temuan ini mengindikasikan bahwa algoritma tersebut efektif dalam melakukan analisis sentimen terhadap komentar di media sosial. Penelitian lanjutan disarankan untuk menggunakan metode lain guna membandingkan hasil dan meningkatkan performa analisis.
2023	[4]	Penelitian ini bertujuan untuk menganalisis sentimen masyarakat terhadap sistem e-Tilang dengan menggunakan komentar dari platform YouTube. Pendekatan yang digunakan melibatkan algoritma *Naïve Bayes* untuk mengklasifikasikan sentimen tanpa memerlukan pemodelan statistik yang kompleks.	Hasil penelitian menunjukkan bahwa dari 500 komentar yang dianalisis, algoritma *Naïve Bayes* berhasil mencapai tingkat akurasi sebesar 79,44%. Proses *preprocessing* dilakukan untuk mengurangi *noise* dalam data, dan penerapan seleksi fitur terbukti mampu mengoptimalkan performa analisis sentimen terhadap sistem e-Tilang.
2024	[6]	Penelitian ini bertujuan untuk menganalisis sentimen pengguna YouTube terhadap Konferensi Tingkat Tinggi (KTT) G20 tahun 2022 dengan memanfaatkan komentar yang diperoleh melalui YouTube API. Analisis dilakukan dengan pendekatan text preprocessing, pelabelan sentimen menggunakan metode VADER, serta klasifikasi sentimen menggunakan algoritma *Naïve Bayes*.	Dari 19.215 komentar YouTube terkait KTT G20 2022, ditemukan 51,5% komentar positif dan 48,5% negatif. Dengan *Naïve Bayes* dan *5-fold Cross Validation*, diperoleh akurasi 77%, *F1-score* 76%, *precision* 85%, dan *recall* 69%. Hasil ini menunjukkan bahwa *Naïve Bayes* efektif untuk analisis sentimen, meskipun disarankan peningkatan pada tahap *preprocessing* teks.
2024	[7]	Penelitian ini bertujuan untuk menganalisis sentimen masyarakat Indonesia terhadap mobil listrik melalui komentar di YouTube dengan menggunakan algoritma *Naïve Bayes* dan pendekatan *Knowledge Discovery in Databases* (KDD). Tujuan lainnya adalah mengevaluasi efektivitas teknik SMOTE *Upsampling* dalam meningkatkan performa klasifikasi sentimen pada data yang tidak seimbang.	Hasil penelitian menunjukkan bahwa rendahnya pemahaman publik menjadi faktor dominan munculnya sentimen negatif terhadap mobil listrik. Penerapan teknik SMOTE *Upsampling* berhasil meningkatkan akurasi model dari 50,70% menjadi 70,69%, meskipun nilai presisi dan *recall* masih memerlukan perbaikan. Temuan ini mengindikasikan bahwa algoritma *Naïve Bayes* memiliki potensi dalam analisis sentimen, namun disarankan untuk memperluas jumlah data, mencoba algoritma lain, dan melakukan pengujian lanjutan guna meningkatkan akurasi serta mendukung penerimaan mobil listrik di Indonesia.

Penelitian ini bertujuan untuk menerapkan metode analisis sentimen menggunakan algoritma *Naive Bayes* pada komentar YouTube yang membahas isu dugaan bisnis gelap antara dokter dan perusahaan farmasi. Dengan menggunakan algoritma ini, dapat diketahui bagaimana opini publik terbentuk, seberapa besar sentimen negatif yang muncul, serta seberapa efektif *Naive Bayes* dalam mengklasifikasikan sentimen pada data dengan karakteristik bahasa yang beragam seperti di media sosial. Hasil dari penelitian ini juga dapat menjadi kontribusi bagi upaya pemantauan opini publik terhadap isu-isu etika di bidang kesehatan melalui pendekatan teknologi informasi.

2. METODE PENELITIAN

2.1 Data Penelitian

Data yang digunakan dalam penelitian ini diperoleh dengan memanfaatkan proses *crawling* yaitu proses pengambilan data yang tersedia pada *social media* YouTube dengan memindahkan informasi ataupun data yang didapatkan berdasarkan perintah tertentu ke *file local* di dalam komputer. Pada penelitian ini teknik *crawling* digunakan pada media sosial YouTube, seperti yang ditunjukan pada Gambar 2.



Gambar 2. *Crawling* Data

2.2 Algoritma *Naïve Bayes*

Naive Bayes adalah algoritma yang digunakan untuk mengklasifikasikan suatu variabel berdasarkan pendekatan probabilitas dan statistika [6], [8], [9], [10]. Algoritma ini mengasumsikan bahwa setiap variabel input saling independen dalam memengaruhi hasil klasifikasi, yang dikenal dengan istilah *class conditional independence*. Prinsip kerja *Naive Bayes* didasarkan pada *Teorema Bayes* mengenai probabilitas bersyarat. Notasi matematis dari algoritma ini dapat dituliskan dalam bentuk persamaan (1).

$$P(A|B) = \frac{\mathrm{P(B|A)}\ x\ P(A)}{\mathrm{P(B)}} \qquad (1)$$

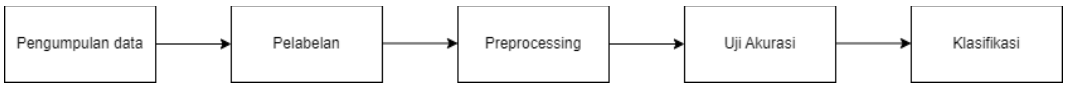
Metode *Naive Bayes* merupakan algoritma yang digunakan untuk menentukan nilai probabilitas tertinggi berdasarkan pendekatan statistik *bayes* sederhana, serta mengklasifikasikan data uji ke dalam kategori yang paling sesuai. Algoritma ini mampu memproses data dalam jumlah besar dengan tingkat akurasi yang cukup tinggi. Adapun bentuk persamaan dari *Naive Bayes* telah disesuaikan sesuai pada Persamaan (2).

$$P(Wk|Ci) = \frac{|\mathrm{ni} + 1|}{|\mathrm{n} + \mathrm{kosakata}|} \qquad (2)$$

Nilai ini menunjukkan frekuensi kemunculan kata Wk dalam dokumen yang termasuk ke dalam kategori Ci, sementara n merepresentasikan total jumlah kata dalam seluruh dokumen dengan kategori Ci. Adapun |kosakata| mengacu pada jumlah seluruh kata unik yang terdapat dalam data pelatihan.

2.3 Penerapan Metode

Sebuah perancangan aplikasi analisis sentimen yang menggunakan metode *Naïve Bayes* perlu adanya beberapa tahapan alur kerja penelitian yang sedang dilakukan agar penelitian ini menjadi terstruktur dan sesuai tujuan. Alur kerja tersebut dapat dilihat pada Gambar 3.

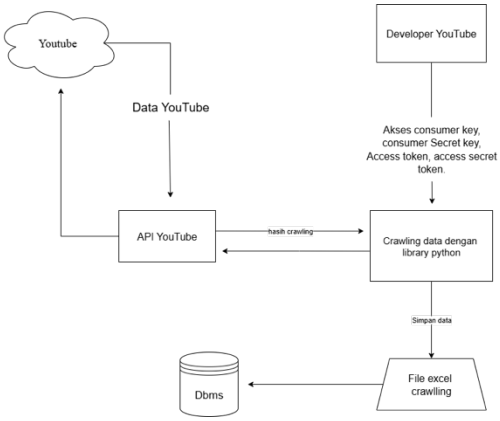


Gambar 3. Tahapan Metode Penelitian

Pada Gambar 3 tahapan awal penelitian dimulai dengan pengumpulan data bersumber dari YouTube dengan Teknik *crawling*. Data ini yang bisa kita sebut dengan *raw data,* disimpan kedalam *file excel*. Setelah itu diimpor ke *database* melalui aplikasi analisis sentiment berbasis web untuk dilakukan proses penggantian jenis huruf, pembersihan *noise*, mengganti dan menghilangkan kata yang tidak layak diproses lebih lanjut yaitu memberi label kelas *sentiment*. Pelabelan ini dilakukan secara manual. Data yang sudah terlabel akan dilakukan pembobotan *TF-IDF* dan dipecah yang salah satunya menjadi data latih untuk dilakukan pemodelan data. Data yang sudah dimodelkan, dilanjut dengan melakukan proses klasifikasi dan pengajuan atau perhitungan keakuratan hasil seperti akurasi

2.3.1 Pengumpulan Data

Pengumpulan data pada penelitian ini dilakukan dengan teknik *crawling* data menggunakan Bahasa pemrograman *python* yang memanfaatkan *API* Youtube. *API* Youtube ini dapat diakses dengan cara mendaftar dan membuat proyek pada akun *developer* Youtube untuk dapat mendapatkan *consumer key, consumer secret key, access token, access token secret* sebagi syarat pengambilan data Youtube. Proses pengumpulan data ditunjukkan pada Gambar 4.



Gambar 4. Metode Pengumpulan Data

2.3.2 Pelabelan

Proses pelabelan data menjadi tahap penting dalam analisis sentimen karena berfungsi menentukan kategori pada setiap komentar. Dalam penelitian ini, data yang digunakan berupa 250 komentar dari kanal YouTube MALAKA terkait isu bisnis gelap dokter dan perusahaan farmasi. Komentar diberi label secara manual (manual annotation) dengan mempertimbangkan makna dan konteks isinya. Dua kategori sentimen yang digunakan adalah positif, jika komentar berisi dukungan atau apresiasi, dan negatif, jika mengandung kritik, kekecewaan, atau penolakan. Untuk menjaga akurasi, proses anotasi dilakukan lebih dari satu pihak sehingga dapat meminimalisasi subjektivitas dan memastikan konsistensi sebelum data digunakan pada tahap preprocessing dan pelatihan model Naive Bayes.

2.3.3 Klasifikasi Multinominal *Naïve Bayes*

Tahap klasifikasi menggunakan *Naive Bayes* merupakan tahap pelatihan terhadap dokumen yang sudah dilakukan *preprocessing* dan pelabelan untuk memperoleh hasil sentimen proses klasifikasi memerlukan data pelabelan manual dari data latih dengan tujuan untuk membuat model terdapat dua langkah dalam proses klasifikasi sebagai berikut:

a. Membangun suatu model dengan menganalisis data *training*.
b. Melakukan klasifikasi, dimana model yang telah dihasilkan digunakan untuk melakukan klasifikasi terhadap data yang belum diketahui labelnya.

2.4 Rancangan Pengujian

Adapun rancangan pengujian pada sistem analisis *sentiment* ialah sebagai berikut:

2.4.1 Rancangan Pengujian Metode

Pengujian dilakukan untuk mengetahui *sentiment* dari media sosial YouTube. pada penelitian ini digunakan metode pengujian dengan implementasi *system* aplikasi berupa *website* yang dibangun dengan memasukan *dataset* kedalam *database system* tersebut kemudia *system* dapat melakukan *labelling, preprocessing* dan *classification* dengan menggunakan evaluasi perhitungan klasifikasi yaitu persentase akurasi.

a. Akurasi

Akurasi adalah Tingkat kedekatan antara nilai prediksi data dan juga nilai sebenarnya. Akurasi nilai ditentukan dengan membandingkan data yang terklasifikasi benar dengan keseluruhan data. Pencapaian nilai akurasi dilihat dari persamaan berikut. Tingkatan kedekatan antara prediksi nilai suatu data dengan nilai *actual*. Nilai akurasi didapat dari perbandingan melalui data yang terklasifikasi benar dengan keseluruhan data. Perolehan hasil akurasi dapat dilihat pada persamaan (3).

$$\boldsymbol{Akurasi} = \frac{\boldsymbol{TP + FN}}{\boldsymbol{TP + TN + FP + FN}} \quad (3)$$

Keterangan:

1. *True Positive* (TP)
   True positive adalah data positif yang benar diprediksi.
2. *True Negative* (TN)
   True Negative adalah data *negative* yang benar diprediksi.
3. *False Positive* (FP)
   False Positive adalah data *negative* namun diprediksikan sebagai data positif.
4. *False Negative* (FN)
   False Negative adalah data *positive* namun diprediksikan sebagai data *negative*.

b. Presisi

Presisi merupakan ketepatan antara data *actual* yang diperlukan pada jawaban yang diberi. Dalam penelitian ini dilakukan dengan akurasi rata-rata. Nilai tersebut dapat dihitung dengan menggunakan persamaan (4).

$$\boldsymbol{Presisi} = \frac{\textbf{Presisi Positif} + \textbf{Presisi Negatif}}{\boldsymbol{2}} \quad (4)$$

c. *Recall*

Recall merupakan tingkat kesuksesan sistem saat mengambil informasi. Dalam penelitian ini dilakukan dengan menghitung nilai *return* rata-rata. Setelah itu ditentukan nilai dengan menggunakan persamaan (5)

$$\boldsymbol{Recall} = \frac{\boldsymbol{Recall}\ \textbf{Positif} + \boldsymbol{Recall}\ \textbf{Negatif}}{\boldsymbol{2}} \quad (5)$$

2.4.2 Rancangan Pengujian Sistem

Pengujian dilakukan dengan menggunakan metode *Black Box Testing*. *Black box testing* adalah pengujian yang dilakukan dengan tahap pengamatan, pada hasil eksekusi melalui beberapa langkah fungsional dalam perangkat lunak. Pada tahap ini akan dijelaskan mengenai proses alur pada system yang akan dibuat dari *import* datam *labelling* hingga menghasilkan hasil klasifikasi evakuasi nilai akurasi.

3. HASIL DAN PEMBAHASAN

3.1 Implementasi Metode

Pada bagian implementasi metode tentang analisis sentimen Masyarakat terhadap bisnis gelap dokterdan perusahaan farmasi dilakukan empat tahapan utama. Yaitu tahapan tersebut diproses secara berurutan, tahapan utama yang dimaksud meliputi pengumpulan data, pelabelan, *preprocessing*, *Naïve Bayes*.

3.1.1 Pengumpulan data

Pada tahap awal, yaitu pengumpulan dataset, dataset dikumpulkan melalui proses *crawling* yang berlangsung dari 13 Juni 2025 hingga 20 Juni 2025. Hasil *crawling* disimpan dalam format CSV, dan kemudian dataset ini

akan diimpor satu per satu ke sistem *website* ke sistem *database* untuk diproses lebih lanjut. Tabel di bawah menunjukkan sampel dataset analisis *sentiment* untuk isu bisnis gelap dan perusahaan farmasi.

Tabel 2. Pengumpulan data

No.	*Username*	Komentar
1.	@rizkiadisaputra5971	Gibrann gibrann ampun dah
2.	@Tania_Mom	DURASI KURANG SAAATTTGGGHHH
3.	@LexeLite	Audionya tolong diperhatiin, pas ngobrol suara kecil pas ketawa suara kenceng
4.	@erlanggasamudra3538	Suaranya terlalu kecil, tidak terlalu kedengaran walau pakai earphone juga
5.	@k-projek2	setuju
6.	@meatheadbali7558	Sebagai mantan medref saya sangat tau harga asli obat dan perjanjian perusahaan dengan dokter ðŸ˜, Setiap kunjungan selalu bawain makanan atau minuman dokternya biar mau pake obat dari perusahaan saya ðŸ˜,ðŸ˜
7.	@daniar8447	Menurut gw sebagai orang awam, kayaknya seakan akan yang diusahakan beberapa dokter atau Fk yang barier awalnya kuantiti daripada kuality itu yang dipikirkan adalah UANG deh. Soalnya kalo aja lebih mikir kepentingan utama, mereka akan lebih memilih quality daripada quantity.
8.	@RoisWorksTech	saran aja abang editorku, sound nya di limit, kadang kaget kalau ferry tiba tiba ketawa coy ðŸ˜,
9.	@IlhamSaputra-o2r	,latar y terlalu gelap
...	.....	.....
250	@ine567	Saya nakes awal

3.1.2 Perlabelan

Setelah pengumpulan data dari hasil *crawling*, tahap pelabelan dilakukan secara manual. Label positif digunakan untuk komentar yang menyatakan pendapat dengan cara yang positif dan mendukung setiap capres. Label negatif digunakan untuk komentar yang menyatakan pendapat dengan kata-kata kasar, sara, benci, atau ujaran kebencian terhadap dokter dan perusahaan farmasi yang curang.

Tabel 3. Perlabelan

No.	Komentar	Label
1.	Sebagai mantan medref saya sangat tau harga asli obat dan perjanjian perusahaan	Negatif
2.	Terimakasih dok Tirta untuk speak up mewakili kita yang blm mampu bersuara dok. Mantap untuk bang Ferry dan bang Coki untuk pertanyaan kritisnya tentang dunia kedokteran. Sukses sehat selalu semua	Positif
3.	Saya alami sendiri, obat mahal dihentikan dokter padahal tidak boleh putus, saya drop sebulan	Negatif
4.	Percaya atau tidak, obat yg digunakan dokter yg paling ampuh trust yg mensugesti. Yaa gak sih?	Positif
5.	Aku pernah kaget waktu diperiksa di rs Jepang... Dari sini gw sadar, arti dokter yang melayani sebenarnya	Positif
...	…	...
250	Dokter pintar vs dokter kurang pintar efek ke pasienya separah itu	Negatif

3.1.3 *Preprocessing*

Setelah melakukan pelabelan, tahap *preprocessing* datang. Komentar yang telah melewati tahap ini akan dibagi menjadi dua bagian, antara lain. Data uji dan latih sendiri adalah sumber pengetahuan untuk proses klarifikasi dan pembangunan pengetahuan.

a. *Case Folding*

Case folding merupakan salah satu tahap awal dalam preprocessing teks yang bertujuan untuk menormalkan data agar lebih konsisten. Pada tahap ini, seluruh huruf dalam teks diubah menjadi huruf kecil atau *lowercase*, sehingga tidak ada perbedaan antara huruf kapital dan huruf nonkapital. Proses ini penting dilakukan untuk menyederhanakan analisis, karena dalam klasifikasi teks, huruf besar dan kecil dianggap memiliki makna yang sama. Dengan adanya *case folding*, potensi redundansi akibat perbedaan penulisan huruf dapat diminimalisasi, sehingga data teks menjadi lebih seragam dan siap diproses pada tahap berikutnya.

Tabel 4. *Case Folding*

No.	Komentar	*Case folding*
1.	Gibrann gibrann ampun dah	gibrann gibrann ampun dah
2.	DURASI KURANG SAAATTTGGGHHH	durasi kurang saaatttggghhh
3.	Audionya tolong diperhatiin, pas ngobrol suara	audio tolong diperhatiin pas ngobrol suara
4.	Suaranya terlalu kecil, tidak terlalu kedengaran walau pakai earphone juga	suaranya terlalu kecil, tidak terlalu kedengaran walau pakai earphone juga
5.	Setuju	setuju

6.	Sebagai mantan medref saya sangat tau harga asli obat dan perjanjian perusahaan dengan dokter ðŸ˜, Setiap kunjungan selalu bawain makanan atau minuman	sebagai mantan medref saya sangat tau harga asli obat dan perjanjian perusahaan dengan dokter ðŸ˜, Setiap kunjungan selalu bawain makanan atau minuman
7.	Menurut gw sebagai orang awam, kayaknya seakan akan yang diusahakan beberapa dokter atau Fk yang barier awalnya kuantiti daripada kuality itu yang dipikirkan adalah UANG deh. Soalnya kalo aja lebih mikir kepentingan utama, mereka akan lebih memilih quality daripada quantity.	menurut gw sebagai orang awam, kayaknya seakan akan yang diusahakan beberapa dokter atau fk yang barier awalnya kuantiti daripada kuality itu yang dipikirkan adalah uang deh. soalnya kalo aja lebih mikir kepentingan utama, mereka akan lebih memilih quality daripada quantity.
8.	saran aja abang editorku, sound nya di limit, kadang kaget kalau ferry tiba tiba ketawa coy ðŸ˜,	saran aja abang editorku, sound nya di limit, kadang kaget kalau ferry tiba tiba ketawa coy ðŸ˜,
9.	latar y terlalu gelap	latar y terlalu gelap
...	...	...
250	itu yang belakang dr. Tirta siapa bang ?	itu yang belakang dr. tirta siapa bang ?

b. *Stopword*

Stopword adalah kumpulan kata-kata umum dalam suatu bahasa yang sering muncul namun biasanya dihapus dalam tahap pemrosesan teks, karena dianggap tidak memberikan informasi penting atau makna yang signifikan terhadap proses analisis.

Tabel 5. *Stopword*

No.	Komentar	*Stopword*
1.	Gibrann gibrann ampun dah	Gibrann Gibrann
2.	DURASI KURANG SAAATTTGGGHHH	DURASI SAAATTTGGGHHH
3.	Audionya tolong diperhatiin, pas ngobrol suara	Audionya diperhatiin ngobrol suara
4.	Suaranya terlalu kecil, tidak terlalu kedengaran walau pakai earphone juga	Suaranya kecil kedengaran pakai earphone
5.	Setuju	Setuju
6.	Sebagai mantan medref saya sangat tau harga asli obat dan perjanjian perusahaan dengan dokter ðŸ˜, Setiap kunjungan selalu bawain makanan atau minuman	mantan medref tau harga asli obat perjanjian perusahaan dokter kunjungan bawain makanan minuman
7.	Menurut gw sebagai orang awam, kayaknya seakan akan yang diusahakan beberapa dokter atau Fk yang barier awalnya kuantiti daripada kuality itu yang dipikirkan adalah UANG deh. Soalnya kalo aja lebih mikir kepentingan utama, mereka akan lebih memilih quality daripada quantity.	gw orang awam dokter Fk barier awalnya kuantiti kuality dipikirkan UANG mikir kepentingan utama memilih quality quantity
8.	saran aja abang editorku, sound nya di limit, kadang kaget kalau ferry tiba tiba ketawa coy ðŸ˜,	saran abang editorku sound limit kaget ferry ketawa
9.	latar y terlalu gelap	latar gelap
10	itu yang belakang dr. Tirta siapa bang ?	belakang dr. Tirta bang

c. *Filtering*

Filtering merupakan tahap pembersihan data teks dengan cara menghapus karakter khusus, angka, dan simbol yang dianggap tidak relevan sehingga teks menjadi lebih bersih dan siap digunakan untuk analisis lebih lanjut.

Tabel 6. *filtering*

No.	Komentar	*Filtering*
1.	Gibrann gibrann ampun dah	Gibrann gibrann ampun dah
2.	DURASI KURANG SAAATTTGGGHHH	DURASI KURANG SAAATTTGGGHHH
3.	Audionya tolong diperhatiin, pas ngobrol suara	Audionya tolong diperhatiin pas ngobrol suara
4.	Suaranya terlalu kecil, tidak terlalu kedengaran walau pakai earphone juga	Suaranya terlalu kecil tidak terlalu kedengaran walau pakai earphone juga
5.	Setuju	Setuju
6.	Sebagai mantan medref saya sangat tau harga asli obat dan perjanjian perusahaan dengan dokter ðŸ˜, Setiap kunjungan selalu bawain makanan atau minuman	Sebagai mantan medref saya sangat tau harga asli obat dan perjanjian perusahaan dengan dokter Setiap kunjungan selalu bawain makanan atau minuman
7.	Menurut gw sebagai orang awam, kayaknya seakan akan yang diusahakan beberapa dokter atau Fk yang barier awalnya kuantiti daripada kuality itu yang dipikirkan adalah UANG deh. Soalnya kalo aja lebih	Menurut gw sebagai orang awam kayaknya seakan akan yang diusahakan beberapa dokter atau Fk yang barier awalnya kuantiti daripada kuality itu yang dipikirkan adalah UANG deh Soalnya kalo aja lebih mikir

	mikir kepentingan utama, mereka akan lebih memilih quality daripada quantity.	kepentingan utama mereka akan lebih memilih quality daripada quantity
8.	saran aja abang editorku, sound nya di limit, kadang kaget kalau ferry tiba tiba ketawa coy ðŸ˜,	saran aja abang editorku sound nya di limit kadang kaget kalau ferry tiba tiba ketawa coy
9.	latar y terlalu gelap	latar y terlalu gelap
10.	itu yang belakang dr. Tirta siapa bang ?	itu yang belakang dr Tirta siapa bang

Setelah *preprocessing*, dilakukan proses ekstraksi fitur dengan menghitung frekuensi kemunculan kata menggunakan metode *Term Frequency* (TF). Fitur ini menjadi input utama dalam proses klasifikasi. Selanjutnya adalah tahap klasifikasi menggunakan algoritma *Naïve Bayes*. Algoritma ini bekerja berdasarkan probabilitas, di mana setiap komentar diklasifikasikan ke dalam tiga kategori sentimen: positif dan negatif, dengan mengacu pada nilai probabilitas tertinggi dari masing-masing kelas

3.1.4 Pemodelan *Naïve Bayes*

Setelah selesai melakukan pelabelan dan *preprocessing* data, langkah selanjutnya adalah membagi data menjadi data latih dan data uji, kemudian dilakukan eksperimen dengan data latih menggunakan metode 10 kali percobaan.

3.2 Pengujian Metode

Pengujian dilakukan untuk mengetahui performa algoritma *Naïve Bayes* dalam mengklasifikasikan komentar YouTube berdasarkan sentimen. Data dibagi menggunakan teknik *stratified sampling*, yang memastikan bahwa distribusi sentimen dalam data latih dan data uji tetap proporsional.

Tiga rasio pembagian data digunakan untuk menguji performa model, yaitu 70:30, 80:20, dan 90:10. Pengujian dilakukan sebanyak 10 kali pada setiap rasio, dan hasil yang ditampilkan merupakan rata-rata dari seluruh pengujian.

Evaluasi performa model dilakukan menggunakan empat metrik utama, yaitu:

a. Akurasi: tingkat keseluruhan prediksi yang benar terhadap seluruh data uji.
b. Presisi: ketepatan model dalam mengklasifikasikan komentar ke dalam kelas tertentu.
c. *Recall*: sejauh mana model mampu menangkap semua data yang relevan untuk suatu kelas.
d. *F1-Score*: rata-rata harmonis dari presisi dan *recall*.

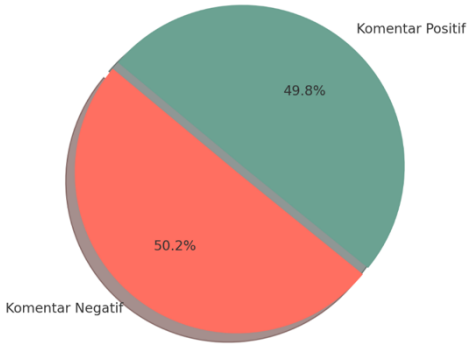
Berikut adalah hasil pengujian terbaik yang diperoleh pada rasio pembagian data 80:20

Tabel 7. Hasil Pengujian

Metrik	Nilai %
Akurasi	88%
Presisi	88,5%
Recall	90,5%
F1-Score	88%

3.3 Hasil Analisis Sentimen

Setelah melalui proses klasifikasi, diperoleh hasil distribusi sentimen dari 250, komentar YouTube yang dianalisis pada video Malaka *Podcast* bersama dr. Tirta. Hasil klasifikasi sentimen adalah sebagai berikut:



Gambar 5. Grafik hasil analisis sentimen

Berikut adalah grafik pie chart yang menunjukkan distribusi sentimen komentar youtube pada video malaka podcast bersama Dr. Tirta:

a. Komentar Positif (49,6%), ditampilkan dengan warna hijau kebiruan, menunjukkan adanya apresiasi dari sebagian penonton terhadap konten atau penyampaian informasi dalam video. Contohnya : " Aku pernah kaget waktu diperiksa di rs Jepang... Dari sini gw sadar, arti dokter yang melayani sebenarnya."
b. Komentar Negatif (50,0%), ditampilkan dengan warna merah muda, menunjukkan dominasi sentimen negatif. Ini mengindikasikan bahwa banyak penonton merasa kritis atau kecewa terhadap isu yang dibahas. Contohnya : "Gak ada drama, gak ada alesan… lancar kayak aer!."

4. KESIMPULAN

Masalah utama penelitian ini adalah bagaimana mengklasifikasikan sentimen komentar YouTube yang beragam, tidak baku, dan seringkali ambigu terkait isu bisnis gelap antara dokter dan perusahaan farmasi. Melalui penerapan algoritma *Naïve Bayes*, permasalahan tersebut dapat diatasi dengan hasil akurasi tertinggi sebesar 92,6%, menunjukkan efektivitas metode dalam menangani kompleksitas data teks. Hasil analisis juga memperlihatkan bahwa sentimen positif lebih dominan, yang berarti sebagian besar pengguna cenderung menerima atau tidak menolak isu yang diangkat. Dengan demikian, penelitian ini tidak hanya menjawab permasalahan klasifikasi sentimen, tetapi juga memberikan gambaran nyata mengenai persepsi publik yang dapat menjadi bahan pertimbangan bagi pihak terkait dalam merumuskan kebijakan dan strategi komunikasi.

DAFTAR PUSTAKA

[1] Q. A. Xu, V. Chang, and C. Jayne, "*A systematic review of social media-based sentiment analysis: Emerging trends and challenges," Decision Analytics Journal*, vol. 3, p. 100073, Jun. 2022, doi: 10.1016/j.dajour.2022.100073.
[2] H. A. R. Harpizon, R. Kurniawan, Iwan Iskandar, R. Salambue, E. Budianita, and F. Syafria, "Analisis Sentimen Komentar Di YouTube Tentang Ceramah Ustadz Abdul Somad Menggunakan Algoritma Naïve Bayes," *JNKTI (Jurnal Nasional Komputasi dan Teknologi Informasi)*, vol. 5, no. 1, pp. 131–140, 2022.
[3] Z. N. Aulia, G. K. Jati, and I. Santoso, "Analisis Sentimen Tanggapanpublic Mengenai E-Tilang Melalui Media Sosial Youtube Menggunakan Algoritma Naive Bayes," *Jurnal IKRA-ITH Informatika*, vol. 7, no. 2, pp. 150–156, 2023.
[4] A. Karimah, G. Dwilestari, and M. Mulyawan, "Analisis Sentimen Komentar Video Mobil Listrik Di Platform Youtube Dengan Metode Naive Bayes," *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. 8, no. 1, pp. 767–737, 2024, doi: 10.36040/jati.v8i1.8373.
[5] E. Y. S. Sihombing, Tibyani, and B. T. Hanggara, "Pemanfaatan API Youtube dalam Pengembangan Aplikasi Portal Video Penangkaran Kenari untuk Peternak Kenari Berbasis Android," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, vol. 4, no. 7, pp. 2067–2074, 2020, [Online]. Available: http://j-ptiik.ub.ac.id
[6] R. A. Firsttama, A. A. Arifiyanti, and D. S. Y. Kartika, "Analisis Sentimen Komentar Youtube Konferensi Tingkat Tinggi G20 Menggunakan Metode Naive Bayes," *Jurnal Teknologi Dan Sistem Informasi Bisnis*, vol. 6, no. 2, pp. 282–285, Apr. 2024, doi: 10.47233/jteksis.v6i2.1263.
[7] A. Karimah, G. Dwilestari, and M. Mulyawan, "analisis sentimen komentar video mobil listrik di platform youtube dengan metode naive bayes," *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. 8, no. 1, pp. 767–737, Mar. 2024, doi: 10.36040/jati.v8i1.8373.
[8] K. Zerrouki, R. M. Hamou, and A. Rahmoun, "*Sentiment Analysis of Tweets Using Naïve Bayes, KNN, and Decision Tree," ... Sentiment Analysis Across ...,* 2022, [Online]. Available: https://www.igi-*global.com/chapter/sentiment-analysis-of-tweets-using-nave-bayes-knn-and-decision-tree*/308507.
[9] E. Salim and A. Solichin, "analisis sentimen pada media sosial twitter terhadap pelayanan dinas kependudukan dan pencatatan sipil menggunakan algoritma naïve bayes," 2022. [Online]. *Available*: http://jom.fti.budiluhur.ac.id/index.php/IDEALIS/indexEmilSalim|http://jom.fti.budiluhur.ac.id/index.php/IDEALIS/index|.
[10] Muhammad Ali Akbar and Achmas Solichin, "Perbandingan Sentimen Ulasan Pengguna Aplikasi Ride-Hailing Gojek dan Grab Menggunakan Algoritma Multinomial Naïve Bayes," *KRESNA: Jurnal Riset dan Pengabdian Masyarakat*, vol. 4, no. 1, pp. 1–11, May 2024, doi: 10.36080/kresna.v4i1.129.