BAB III

METODOLOGI PENELITIAN

3.1 Gambaran Umum Objek Penelitian

Tujuan dari penelitian ini adalah untuk meneliti sikap pengguna media sosial terkait kontan kesehatan mental di TikTok dan bagaimana sikap tersebut memengaruhi kecenderungan pengguna untuk melakukan *self-diagnose*. Sebagai fokus penelitian, konten kesehatan TikTok akan diteliti untuk menentukan sikap yang diungkapkan di dalamnya. Algoritma *Extreme Gradient Boosting* (XGBoost) dan *Support Vector Machine* (SVM) akan digunakan dalam proyek ini untuk menangani data komentar yang berkaitan dengan konten kesehatan mental. Proses secara garis besar yang dilakukan dalam penelitian ini, yaitu:

1. Komentar TikTok tentang kesehatan mental akan dianalisis sentimennya menggunakan metode pelabelan berbasis aturan (*rule-based labeling*). Pada metode ini, komentar dianalisis berdasarkan keberadaan kata kunci tertentu yang telah ditentukan sebelumnya untuk menentukan apakah sentimennya termasuk netral, negatif, atau positif. Misalnya komentar yang mengandung kata yang merepresentasikan perilaku *self-diagnose*, maka diberi label positif, apabila tidak ada representasi perilaku tersebut, maka akan masuk ke label negatif, dan jika apabila tidak cukup kuat untuk dikatakan positif, namun tidak cukup kuat pula untuk dilabeli negatif, maka akan masuk ke dalam pelabelan netral.
2. *Support Vector Machine* (SVM) dan *Extreme Gradient Boosting* (XGBoost) digunakan untuk membangun dan mengevaluasi model klasifikasi sentimen terhadap komentar pengguna TikTok yang membahas konten kesehatan mental. Komentar terlebih dahulu dilabeli menggunakan pendekatan *rule-based* berdasarkan kata kunci tertentu, lalu digunakan sebagai data latih untuk kedua model. SVM dan XGBoost kemudian dibandingkan untuk melihat performanya dalam mengklasifikasikan sentimen komentar ke dalam tiga kategori: positif, negatif, dan netral.

Objek dalam penelitian ini adalah konten kesehatan mental pada TikTok. Data kesehatan mental ini relevan karena TikTok memiliki algoritma yang menyesuaikan tampilan konten sesuai dengan minat penggunanya. Hal ini mempengaruhi jenis konten yang sering dilihat oleh pengguna, termasuk konten berkaitan dengan kesehatan mental. Penelitian ini bertujuan untuk memahami bagaimana konten kesehatan mental di TikTok dapat memengaruhi kecenderungan pengguna untuk mendiagnosis diri mereka sendiri. Selain itu, penelitian ini juga mengamati bagaimana teknik machine learning SVM dan XGBoost dapat digunakan untuk menganalisis sentimen terkait perilaku pengguna.

Data yang digunakan dalam penelitian ini diperoleh dari hasil *scrapping* komentar pada *platform* TikTok, yang selanjutnya disusun ke dalam beberapa kolom dengan struktur yang sistematis. Untuk memberikan pemahaman yang menyeluruh terhadap struktur data, informasi ini dibagi ke dalam tiga bagian utama yang direpresentasikan melalui Tabel 3.1, 3.2, dan 3.3. Setiap bagian memuat atribut-atribut yang memiliki peranan penting dalam proses analisis yang akan dilakukan pada tahapan berikutnya. Berikut ini merupakan data yang digunakan pada penelitian ini.

Tabel 3. 1 5 Kolom Utama pada Data Mentah

Type	***Comment Id***	***Parent Comment Id***	***User Id***	***Unique Id***
comment	7395760583705166597		6940503746146681857	3dy_tem4
reply	7395773204788970246	7395760583705166597	6557292877344047106	ftryoung
reply	7395789552818176774	7395760583705166597	6984255366784861186	apocalypse18235
reply	7395827299147498245	7395760583705166597	6550985200321560577	taxicolt96
reply	7395848568756503301	7395760583705166597	7006626925557597186	hyunggg_____

Tabel 3.1 menampilkan kolom-kolom utama dari data komentar yang diambil dari platform TikTok. Beberapa informasi penting yang ditampilkan di antaranya adalah Type (menunjukkan apakah data berupa komentar utama atau balasan), Comment Id, Parent Comment Id (untuk menghubungkan komentar dengan balasannya), User Id, serta Unique Id yang merepresentasikan identitas

unik pengguna. Kolom-kolom ini sangat penting dalam proses identifikasi struktur percakapan, pelacakan interaksi antar pengguna, dan membentuk dasar yang kuat untuk proses pengolahan data lebih lanjut.

Tabel 3. 2 Kolom Lanjutan Data Mentah

Nickname	*Avatar*	*Comment*	*Comment Language*	*Reply Comment Total*
3Dy • Teman	https://p16-sign-va.tiktokcdn.com/tos-maliva-avt-0068/75dfdc...	Gw gngguan kecemasan anxiety 😅😅😅	id	119
Young 어린	https://p16-sign-va.tiktokcdn.com/tos-maliva-avt-0068/17d379ff...	Sama + intorvert parah kalau beradaptasi sm bnyk orng mlmnya deman 😅	id	0
sayang	https://p16-sign-va.tiktokcdn.com/tos-maliva-avt-0068/febd4c...	sama🥺	id	0
Wend	https://p16-sign-sg.tiktokcdn.com/tos-alisg-avt-0068/4437fbce...	Sama🥺	id	0
Hyunggg_____	https://p16-sign-sg.tiktokcdn.com/tos-alisg-avt-0068/2ffb3340ef48ff3d0c5b...	Kukira cmn gua yg anxiety klo mlm jd demam	id	0

Tabel 3.2 menyajikan informasi tambahan yang mendeskripsikan lebih lanjut konteks dari komentar yang diberikan oleh pengguna. Di dalamnya terdapat kolom Nickname (nama tampilan pengguna), Avatar (tautan gambar profil), Comment (isi komentar asli), Comment Language (bahasa yang digunakan), serta Reply Comment Total (jumlah balasan yang diterima oleh komentar tersebut). Informasi ini sangat relevan dalam proses analisis sentimen karena memungkinkan peneliti memahami isi, ekspresi, serta konteks linguistik dari komentar. Selain itu, kolom ini juga mendukung proses anotasi manual atau otomatis dalam tahap pelabelan data.

Tabel 3. 3 5 Kolom Akhir Data Mentah

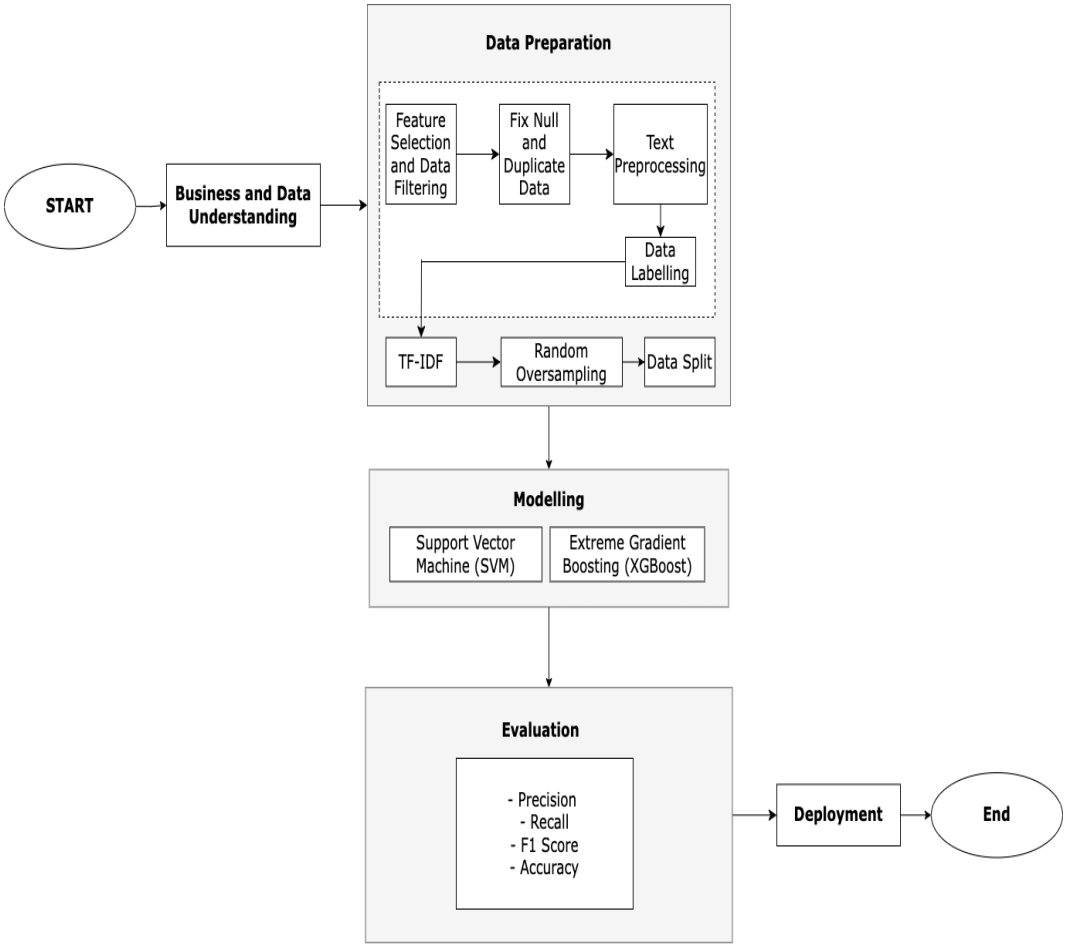
Digg Count	*Create Time*	*Video URL*	*Video Id*	*Scraped At*
1714	2024-07-26T02:08:50.000Z	https://www.tiktok.com/@myuniverse1818/video/7395566830384958725	7395566830384958725	2025-03-10T10:01:52.136Z
97	2024-07-26T02:57:43.000Z	https://www.tiktok.com/@myuniverse1818/video/7395566830384958725	7395566830384958725	2025-03-10T10:01:55.263Z
3	2024-07-26T04:01:08.000Z	https://www.tiktok.com/@myuniverse1818/video/7395566830384958725	7395566830384958725	2025-03-10T10:01:55.263Z
2	2024-07-26T06:27:50.000Z	https://www.tiktok.com/@myuniverse1818/video/7395566830384958725	7395566830384958725	2025-03-10T10:01:55.263Z
4	2024-07-26T07:50:22.000Z	https://www.tiktok.com/@myuniverse1818/video/7395566830384958725	7395566830384958725	2025-03-10T10:01:55.263Z

Tabel 3.3 berisi kumpulan metadata yang penting dalam menghubungkan setiap komentar dengan konteks video sumbernya. Kolom-kolom seperti Digg Count, Create Time, Video URL, Video Id, dan Scraped At menyediakan informasi tambahan yang sangat berguna dalam memperluas pemahaman terhadap dinamika interaksi pengguna. Misalnya, Digg Count menggambarkan jumlah suka pada komentar tertentu, yang dapat menjadi indikator keterlibatan pengguna, sedangkan Create Time menunjukkan kapan komentar tersebut dipublikasikan. Kolom Scraped At memberi informasi waktu aktual pengambilan data, sehingga memungkinkan analisis temporal yang lebih akurat. Dengan menyertakan detail terkait sumber video (Video URL dan Video Id), peneliti dapat menelusuri komentar ke konten asalnya, baik untuk verifikasi maupun untuk studi lanjutan berbasis konteks. Pemahaman menyeluruh terhadap informasi pada tabel ini menjadi sangat krusial dalam merancang strategi pembersihan data, pelabelan, dan eksplorasi pola keterlibatan pengguna terhadap konten kesehatan mental di TikTok.

3.2 Metode Penelitian

3.2.1 Alur Penelitian

Penelitian ini menggunakan pendekatan CRISP-ML sebagai metode utama dalam pelaksanaannya [46]. Metodologi ini dijadikan dasar untuk menyusun tahapan-tahapan penelitian secara sistematis. Gambar 3.3 menyajikan alur kerja penelitian berdasarkan metode tersebut.



Gambar 3. 1 Alur Penelitian

Gambar 3.3 Merupakan representasi alur umum dari penelitian ini, yang memperlihatkan rangkaian tahapan sistematis untuk mencapai tujuan yang telah dirumuskan. Penelitian ini mengadopsi pendekatan *Cross-Industry Standard Process for Machine Learning* (CRISP-ML) sebagai kerangka kerja dalam membangun sistem klasifikasi berbasis *machine learning* terhadap komentar-komentar pada konten TikTok yang berkaitan dengan isu kesehatan mental. Pendekatan ini dipilih karena mampu

mengintegrasikan pemahaman konteks bisnis, pengolahan data, hingga evaluasi model secara menyeluruh. Dalam penelitian ini, metode CRISP-ML diterapkan hanya hingga tahap *deployment*, tanpa melibatkan tahapan *monitoring and maintenance*. Hal ini disesuaikan dengan ruang lingkup dan tujuan akademis dari penelitian, yang tidak menargetkan implementasi sistem secara penuh dalam lingkungan produksi. Karena model belum benar-benar digunakan dalam konteks operasional jangka panjang, maka pemantauan performa secara berkelanjutan tidak termasuk dalam cakupan penelitian ini. Terdapat lima langkah utama yang dilakukan dalam penelitian ini berdasarkan proses CRISP-ML, yaitu:

1) *Business and Data Understanding*

   Tujuan dari penelitian ini adalah untuk mengetahui bagaimana konten kesehatan mental TikTok mempengaruhi kecenderungan pengguna untuk melakukan *self-diagnose*. Tujuan spesifiknya adalah melakukan analisis sentimen dengan menggunakan algoritma SVM dan XGBoost, untuk mengklasifikasikan komentar TikTok menjadi sentimen positif, negatif, atau netral. Komentar dari pengguna TikTok yang dikumpulkan menggunakan prosedur *scrapping* dijadikan sebagai sumber data penelitian pada penelitian ini. Data ini terdiri dari 23.872 komentar yang berkaitan dengan konten kesehatan mental. Tujuan dari tahap ini adalah untuk menyelidiki distribusi, struktur, tren, serta kualitas data. Hal ini termasuk menentukan faktor independen dan variabel dependen.

2) *Data Preparation*

   Pada tahap ini, dilakukan pemrosesan data teks yang diawali dengan penyaringan data, lalu dilanjutkan dengan pembersihan data dari nilai null dan data duplikat, lalu terdapat proses *text mining,* yang meliputi *cleansing, case folding, tokenizing, filtering*, dan *stemming.* Setelah dilakukannya *text preprocessing* tersebut, maka data akan melewati tahap *labelling*, untuk mengklasifikasikan data menjadi positif, negatif, dan netral. Selanjutnya data melalui proses TF-IDF, *random oversampling*,

dan split data. Melalui prosedur ini, data dipastikan bersih dan siap untuk dimasukkan ke dalam model *machine learning*.

3) *Modeling*

Tahapan ini mencakup proses pembangunan model klasifikasi dengan menggunakan algoritma SVM dan XGBoost. Model dibentuk melalui proses pelatihan dengan memanfaatkan karakteristik data yang telah disiapkan sebelumnya. Tujuan utama dari tahap ini adalah mengenali pola sentimen dan indikasi *self-diagnose* dalam komentar-komentar yang dianalisis.

4) *Evaluation*

Tahap evaluasi difokuskan pada pengukuran performa model dengan menggunakan metrik seperti akurasi, presisi, *recall*, dan *f1-score*. Evaluasi ini bertujuan untuk menilai apakah model mampu menjalankan tugas klasifikasi secara efektif. Dengan begitu, hanya model yang memenuhi standar kinerja yang akan dipertimbangkan dalam tahap implementasi.

5) *Deployment*

Pada tahap *deployment*, penelitian ini menggunakan Streamlit sebagai alat utama dalam implementasinya. Streamlit merupakan *framework* berbasis Python yang dirancang untuk membangun antarmuka aplikasi web. Dengan memanfaatkan *framework* ini, proses penyajian hasil model dilakukan secara interaktif dan lebih mudah diakses.

Dalam proses pengembangan model *machine learning* dan analisis data, terdapat beberapa metodologi yang dapat digunakan, di antaranya CRISP-ML (*Cross Industry Standard Process for Machine Learning*) dan CRISP-DM (*Cross Industry Standard Process for Data Mining*). Meskipun keduanya memiliki struktur proses yang serupa, terdapat beberapa perbedaan penting, terutama pada penekanan aspek kualitas dan penerapan berkelanjutan dalam CRISP-ML. Berikut ini disajikan tabel perbandingan antara tahapan dalam metode CRISP-ML dan CRISP-DM:

Tabel 3. 4 Perbandingan CRISP-ML(Q) dan CRISP-DM

Aspek	**CRISP-ML(Q)** [46]	**CRISP-DM** [71]
Tujuan Utama	Menekankan pada pengembangan dan pengujian sistem machine learning secara menyeluruh.	Difokuskan pada eksplorasi data untuk mendukung pengambilan keputusan berbasis data.
Struktur Proses	Menggunakan enam tahapan utama dengan fokus kuat pada rekayasa model dan data.	Memiliki enam tahap yang berurutan namun dapat diulang jika dibutuhkan
Penggabungan Bisnis dan Data	Menggabungkan konteks bisnis dan data dalam satu proses untuk efisiensi.	Memisahkan pemahaman data dan bisnis untuk kejelasan struktur kerja.
Proses Evaluasi	Menyediakan evaluasi model secara terus menerus dalam proses pembangunan.	Evaluasi hanya dilakukan setelah model dikembangkan.
Fleksibilitas Model	Sangat fleksibel dan mendukung perbandingan serta iterasi antar model.	Kurang mendukung eksperimen yang kompleks dalam pengembangan model,

Tabel 3.4 merupakan tabel perbandingan antara metode CRISP-ML(Q) dengan CRISP-DM. Penelitian ini mengadopsi pendekatan metodologi CRISP-ML, dengan merujuk pada struktur proses yang dikembangkan dalam model CRISP-ML(Q) [46]. Namun, aspek jaminan kualitas (*Quality Assurance*) tidak diterapkan secara penuh, karena fokus utama penelitian ini terletak pada tahapan pelabelan otomatis, pembangunan model klasifikasi, serta analisis pola distribusi komentar. Oleh karena itu, penelitian ini hanya menggunakan enam tahapan inti untuk mendukung proses pengembangan dan evaluasi model *machine learning*.

Sebagai pembanding dan penguat validitas pemilihan metodologi, referensi mengenai evolusi CRISP-DM dalam proyek *data science* modern juga digunakan [71]. Meskipun model tersebut telah disesuaikan agar lebih fleksibel dan mendukung pendekatan iteratif, struktur utamanya tetap lebih cocok untuk eksplorasi data yang tidak menekankan pada pembangunan model. Berdasarkan perbandingan dari kedua kerangka tersebut, CRISP-ML dipandang lebih sesuai dengan kebutuhan penelitian ini karena mendukung iterasi, pengujian model secara komparatif (SVM dan XGBoost), serta evaluasi performa secara menyeluruh.

3.2.2 Metode Pengolahan Data

Pada penelitian ini pendekatan *text mining* digunakan dalam metode pemrosesan data untuk memproses data teks dari komentar TikTok. Tujuan dari langkah ini adalah untuk membersihkan dan menyiapkan data teks untuk analisis. Terdapat beberapa langkah pada tahap ini, yaitu [33]:

1) *Cleansing*

   Tahap ini dilakukan dengan cara menghapus elemen-elemen yang tidak relevan dari data teks, seperti emotikon, *hashtag* (#), nama pengguna (*@username*), URL, serta alamat situs web. Penghapusan ini bertujuan untuk mengurangi gangguan (noise) yang dapat menghambat kinerja algoritma saat menganalisis sentimen. Dengan demikian, data yang digunakan menjadi lebih bersih dan terfokus pada informasi yang benar-benar dibutuhkan untuk pemodelan.

2) *Case Folding*

   Langkah ini membantu menjaga konsistensi data dengan mengubah seluruh teks menjadi huruf kecil. Langkah ini juga menghilangkan elemen seperti angka dan tanda baca yang tidak relevan dengan analisis. Langkah ini berguna untuk mengurangi redundansi, seperti kata "TIDAK" dan "tidak" akan dianggang sebagai satu bentuk yang sama.

3) *Tokenizing*

   Pada langkah ini, komentar dipecah menjadi bagian kecil berupa kata atau token. Komenter seperti "Konten ini bagus!" akan dipecah menjadi ["konten", "ini", "bagus"]. Langkah ini berguna untuk mempermudah analisis kata per kata untuk memahami sentimen pengguna terhadap konten kesehatan mental di TikTok.

4) *Filtering*

   Pada tahapan ini, dilakukan proses penyaringan terhadap kata-kata yang termasuk dalam daftar *stopwords*, seperti "dan", "di", atau

"yang", yang dianggap tidak memberikan makna penting dalam analisis. Kata-kata tersebut dihilangkan agar tidak membebani proses klasifikasi. Tujuannya adalah agar model lebih mudah menangkap makna utama dari setiap komentar yang dianalisis.

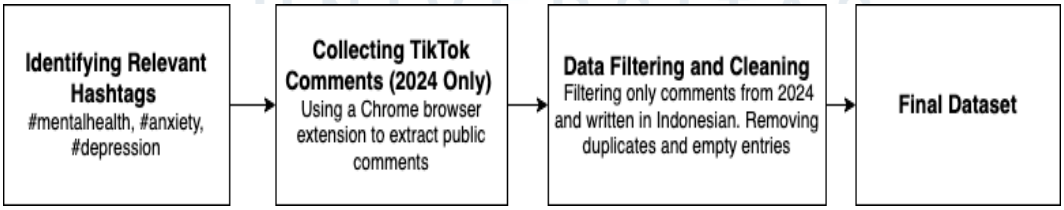
5) *Stemming*

Tahap *stemming* bertujuan untuk mengubah kata-kata berimbuhan menjadi bentuk dasarnya. Sebagai contoh, kata seperti "mendiagnosis" akan dikonversi menjadi "diagnosis" agar lebih seragam dalam representasi datanya. Hal ini penting agar algoritma tidak menganggap dua kata dengan makna serupa sebagai entitas yang berbeda.

Setelah melalui tahapan pra-pemrosesan, data kemudian dianalisis menggunakan algoritma SVM dan XGBoost. Kedua algoritma ini digunakan untuk mengklasifikasikan sentimen yang terkandung dalam konten kesehatan mental di TikTok. Sentimen tersebut dikelompokkan ke dalam tiga kategori utama, yaitu positif, negatif, dan netral.

3.3 Teknik Pengumpulan Data

Dalam penelitian ini, proses pengumpulan data dilakukan secara terstruktur untuk memastikan bahwa data yang digunakan relevan dan berkualitas. Setiap tahapan dilakukan secara sistematis, dimulai dari penentuan tagar yang sesuai hingga penyaringan komentar berdasarkan kriteria tertentu. Untuk itu, dirancang alur proses yang menjelaskan langkah-langkah teknis yang dilakukan selama proses pengumpulan data berlangsung.



Gambar 3. 2 Bagan Tahapan Pengumpulan Data

Gambar 3.4 merupakan bagan tahapan dari berbagai proses yang dilakukan dalam pengumpulan data. Pengumpulan data dalam penelitian ini dilakukan secara otomatis dengan memanfaatkan ekstensi *browser* (*extension*) yang dapat

digunakan melalui Google Chrome. Ekstensi ini memungkinkan peneliti untuk mengambil komentar dari video TikTok tanpa harus membuat kode pemrograman secara manual. Komentar-komentar tersebut dikumpulkan dari video yang diunggah pada tahun 2024, karena pada tahun tersebut isu kesehatan mental menjadi perhatian besar di tengah masyarakat Indonesia. Berdasarkan data dari Asia Care Survey 2024 yang dimuat dalam latar belakang penelitian, berbagai gangguan seperti stres, kecemasan, dan depresi menjadi bentuk kekhawatiran utama. Oleh sebab itu, pemilihan data dari tahun tersebut dinilai paling tepat untuk menggambarkan dinamika konten kesehatan mental dan respons pengguna TikTok secara aktual.

Dalam penelitian ini, konten video TikTok yang dijadikan sumber data diperoleh dengan menggunakan tiga tagar utama, yaitu *#mentalhealth, #anxiety*, dan *#depression*. Ketiga tagar tersebut dipilih karena secara umum mewakili tema kesehatan mental dan sering digunakan dalam video yang membahas atau menarasikan gejala psikologis seperti kecemasan berlebihan, depresi, dan gangguan emosional lainnya. Meskipun komentar yang dianalisis berbahasa Indonesia, penggunaan tagar dalam bahasa Inggris tetap digunakan karena TikTok merupakan *platform* global, dan algoritma pencariannya lebih banyak menampilkan konten populer yang menggunakan istilah dalam bahasa Inggris.

Proses pencarian dilakukan dengan menggunakan tagar tersebut untuk menemukan video yang berkaitan dengan isu kesehatan mental. Setelah itu, komentar yang terdapat pada video-video terpilih dikumpulkan dan dianalisis lebih lanjut dalam konteks kecenderungan *self-diagnose*. Dengan kata lain, tagar digunakan sebagai sarana penyaringan video, bukan sebagai dasar dalam pengambilan komentar. Hal ini diperkuat oleh temuan sebelumnya yang menunjukkan bahwa video dengan tagar seperti *#anxiety* dan *#depression* cenderung menampilkan pengalaman pribadi pengguna dan menarik keterlibatan tinggi, namun juga berisiko menimbulkan pemahaman keliru terkait gangguan mental apabila tidak disertai edukasi profesional [18]. Oleh karena itu, pembatasan pemilihan data melalui ketiga tagar tersebut dinilai relevan dan mendukung tujuan utama penelitian ini

Data yang dikumpulkan berasal dari video TikTok yang bersifat publik, sehingga dapat diakses tanpa memerlukan *login*. Setelah terkumpul, data tersebut melalui tahap penyaringan yang mencakup penghapusan komentar duplikat, komentar kosong, serta komentar yang tidak menggunakan bahasa Indonesia. Selain itu, hanya komentar yang ditulis selama tahun 2024 yang dipertahankan, untuk memastikan bahwa analisis dilakukan terhadap respons relevan. Kumpulan data bersih inilah yang kemudian dijadikan dasar dalam proses analisis sentimen serta prediksi kecenderungan *self-diagnose* pada tahap selanjutnya.

Perlu ditegaskan bahwa komentar yang dikumpulkan dalam penelitian ini bukan seluruhnya merupakan komentar yang mengandung *self-diagnose*. Komentar yang diambil berasal dari video-video bertema kesehatan mental, namun sifatnya *random* (acak) berdasarkan hasil *scraping* otomatis, bukan berdasarkan isi komentarnya sejak awal. Setelah data dikumpulkan dan dibersihkan, proses pelabelan dilakukan untuk mengidentifikasi komentar mana yang menunjukkan indikasi *self-diagnose*, lalu dikategorikan ke dalam tiga kelas sentimen, yaitu positif, netral, dan negatif. Pelabelan ini dilakukan secara otomatis menggunakan pendekatan *rule-based* berbasis kata kunci dan pola kalimat tertentu, kemudian divalidasi oleh pakar dengan latar belakang pendidikan psikologi guna memastikan keakuratan hasil klasifikasi. Dengan demikian, analisis dalam penelitian ini diawali dari kumpulan komentar umum, yang kemudian dikaji dan dipilah berdasarkan konten dan indikasi *self-diagnose*-nya.

3.4 Variabel Penelitian

Pada penelitian ini, variabel dirancang untuk menggambarkan hubungan antara paparan konten kesehatan mental di *platform* TikTok dengan munculnya kecenderungan *self-diagnose* pada pengguna. Setiap variabel disusun berdasarkan fokus penelitian dan disesuaikan dengan jenis data yang dianalisis, sehingga dapat memberikan gambaran yang lebih jelas terhadap pola perilaku pengguna di media sosial. Adapun uraian mengenai jenis variabel yang digunakan akan dijelaskan pada bagian berikut ini.

3.4.1 Variabel Independen

Variabel independen dalam penelitian ini merujuk pada konten-konten bertema kesehatan mental yang banyak dibagikan di platform TikTok. Konten tersebut diperoleh dari video-video yang menyematkan tagar seperti *#mentalhealth*, *#anxiety*, dan *#depression*, kemudian ditelusuri melalui respons pengguna dalam bentuk komentar. Komentar-komentar inilah yang dijadikan sebagai sumber data utama dan dianalisis lebih lanjut untuk melihat kecenderungan sentimen yang muncul. Secara teknis, variabel ini diwujudkan dalam bentuk kolom komentar dalam dataset, yang selanjutnya menjadi bahan untuk penentuan sentimen melalui klasifikasi teks.

3.4.2 Variabel Dependen

Variabel dependen dalam penelitian ini adalah kecenderungan pengguna TikTok untuk melakukan *self-diagnose* berdasarkan isi komentar yang mereka tuliskan. Komentar yang mengandung pengakuan atau pernyataan mengenai kondisi psikologis tertentu, seperti merasa mengalami depresi atau gangguan kecemasan, dikategorikan sebagai indikasi *self-diagnose*. Komentar seperti ini dikelompokkan ke dalam kategori sentimen positif, sedangkan komentar yang tidak menunjukkan ciri-ciri tersebut diklasifikasikan sebagai netral atau negatif. Dalam dataset yang digunakan, kecenderungan *self-diagnose* ini tidak ditandai secara eksplisit melalui kolom khusus, melainkan ditunjukkan secara tidak langsung melalui nilai pada kolom label. Dengan kata lain, label "positif" dalam kolom tersebut mewakili komentar yang mengarah pada perilaku *self-diagnose*.

3.5 Teknik Analisis Data

3.5.1 Analisis Sentimen

Dalam penelitian ini, pendekatan yang digunakan untuk melakukan analisis sentimen adalah *machine learning-based approach*, yaitu pendekatan yang memanfaatkan algoritma pembelajaran mesin untuk

membangun model klasifikasi dari data yang telah diberi label. Komentar TikTok yang berkaitan dengan isu kesehatan mental dianalisis untuk mengetahui kecenderungan *self-diagnose* pengguna. Proses pelabelan data dilakukan secara otomatis menggunakan pendekatan berbasis aturan (*rule-based*), di mana sistem mendeteksi pola kata atau frasa tertentu untuk menetapkan label sentimen (positif, netral, atau negatif). Untuk melakukan analisis sentimen, data yang dikumpulkan dari komentar pengguna TikTok dalam konten kesehatan mental diproses menggunakan metode *text mining* dengan menggunakan algoritma *Support Vector Machine* (SVM) dan XGBoost, komentar diproses terlebih dahulu dan kemudian diklasifikasikan ke dalam tiga kategori sentimen: 1) netral, 2) negatif, 3) positif. Temuan analisis ini akan memberikan gambaran umum mengenai bagaimana pengguna TikTok melihat informasi terkait kesehatan mental.

3.5.2 Pelabelan Data

Proses pelabelan dalam penelitian ini dilakukan secara otomatis menggunakan metode *rule-based*, yakni dengan menerapkan serangkaian aturan berbasis kata kunci dan pola kalimat tertentu yang umum ditemukan dalam komentar terkait perilaku *self-diagnose.* Aturan-aturan tersebut disusun berdasarkan hasil wawancara dan diskusi bersama pakar yang memiliki latar belakang pendidikan psikologi. Pakar yang terlibat telah menyelesaikan pendidikan sarjana (S1) di bidang psikologi dan saat ini sedang menempuh pendidikan profesi psikolog. Melalui proses diskusi tersebut, peneliti memperoleh pemahaman yang lebih mendalam mengenai bagaimana gejala *self-diagnose* diekspresikan dalam bahasa tulis, serta mendapatkan masukan mengenai kata kunci dan konteks kalimat yang relevan.

Pelabelan ini menghasilkan tiga kategori utama, yaitu positif, netral, dan negatif, yang merepresentasikan berbagai tingkat kecenderungan komentar terhadap *self-diagnose.* Setelah pelabelan otomatis dilakukan, seluruh hasilnya divalidasi kembali secara manual oleh pakar untuk memastikan akurasi dan kesesuaian klasifikasi dengan konteks

sebenarnya. Validasi ini penting untuk membedakan komentar yang secara eksplisit mengarah pada *self-diagnose* dari komentar yang bersifat umum, informatif, atau tidak relevan.

3.5.3 Pemisahan Data

Setelah data berhasil diberi label dan melalui tahapan praproses, langkah berikutnya adalah membagi data menjadi dua bagian, yaitu data latih dan data uji. Dalam penelitian ini, digunakan skema pembagian dengan rasio 80:20. Artinya, sebanyak 80% data digunakan untuk proses pelatihan model, sementara 20% sisanya digunakan untuk menguji performa model terhadap data yang belum pernah dilihat sebelumnya.

3.5.4 Evaluasi Kinerja Model

Evaluasi kinerja model akan dilakukan dengan menggunakan metrik evaluasi yang tepat untuk setiap model setelah dilakukannya analisis sentimen menggunakan SVM dan XGBoost. Evaluasi ini menilai kemampuan model dalam mengkategorikan sentimen komentar pengguna TikTok terhadap konten kesehatan mental dengan mengukur akurasi, *precision*, *recall*, dan *F1-score*.

3.5.5 Penggunaan *Tools*

Proses analisis data dalam penelitian ini dilakukan menggunakan Jupyter Notebook, sebuah tools berbasis Python yang umum digunakan dalam bidang data *science* dan *machine learning*. Pemilihannya didasarkan pada kemudahan dalam menulis kode, visualisasi, dan dokumentasi secara interaktif. Meskipun demikian, Jupyter Notebook bukan satu-satunya tools yang dapat digunakan. Terdapat alternatif lain seperti R Studio yang juga sering digunakan untuk analisis data, terutama dalam konteks statistik. Oleh karena itu, tabel 3.5 menyajikan perbandingan antara Jupyter Notebook dan R Studio berdasarkan beberapa aspek penting.

Tabel 3. 5 Perbandingan Jupyter Notebook dan R Studio

Aspek	**Jupyter Notebook** [72]	**R Studio** [73]
Fleksibilitas & Integrasi	Mendukung banyak bahasa (Python, R, Julia); integrasi kode, visualisasi, dan teks.	Khusus untuk R; berbasis GUI melalui R-Shiny untuk analisis diskriminan.
Kolaborasi	Mudah dibagikan lintas *platform*, dokumen interaktif dapat dijalankan ulang.	Akses via web tanpa instalasi; kolaborasi terbatas pada pengguna R.
Pembelajaran	Mendukung eksplorasi data dan pencatatan riset dalam satu *file* yang dapat diulang.	Mudah digunakan oleh *non-programmer* melalui antarmuka visual.
Reproduksibilitas	Seluruh proses riset dapat dijalankan kembali secara konsisten dalam satu *file*.	Reproduksi bergantung pada *script* dan paket R yang digunakan.
Konteks Aplikasi	Ideal untuk simulasi, eksplorasi data, pengajaran, dan kolaborasi multi-bahasa.	Fokus pada analisis diskriminan dengan fitur uji asumsi dan klasifikasi.

Tabel 3.5 merupakan tabel perbandingan antara Jupyter Notebook dengan salah satu *tools* lainnya yaitu R Studio. Salah satu studi mengembangkan paket R yang menggunakan R-Shiny sebagai antarmuka berbasis web untuk analisis diskriminan linear. Aplikasi ini dirancang agar pengguna dapat menjalankan analisis statistik tanpa perlu menulis kode secara langsung dan cukup diakses melalui browser tanpa instalasi tambahan. Proses pengembangan sistem menggunakan pendekatan waterfall, dengan hasil uji coba yang menunjukkan bahwa alat ini mampu menghasilkan analisis yang akurat dan efisien. Beberapa fitur yang tersedia mencakup uji asumsi statistik, estimasi fungsi diskriminan, prediksi klasifikasi, serta penghitungan akurasi [73]. Sementara itu, Jupyter Notebook disebut sebagai platform interaktif yang menggabungkan penulisan kode, visualisasi grafik, dan catatan teks ke dalam satu dokumen yang mudah dibagikan dan dijalankan ulang. Fleksibilitas platform ini dalam mendukung berbagai bahasa pemrograman serta kemudahan dalam mendokumentasikan proses analisis membuatnya sangat relevan untuk proses riset berbasis machine learning [72].

Jika dikaitkan dengan kebutuhan dalam penelitian ini yang mencakup pelabelan otomatis komentar, pembangunan model klasifikasi menggunakan algoritma seperti SVM dan XGBoost, serta eksplorasi distribusi data pengguna, maka Jupyter Notebook menjadi pilihan yang paling tepat. Tidak hanya karena kemampuannya dalam mendukung bahasa Python sebagai bahasa utama yang digunakan, tetapi juga karena fitur interaktifnya memudahkan dokumentasi dan validasi hasil secara langsung. Selain itu, notebook yang dapat dijalankan ulang secara konsisten memungkinkan reproducibility yang tinggi, yang sangat penting ketika proses pelabelan dan pengujian model dilakukan secara bertahap. Dengan mempertimbangkan keseluruhan fitur dan kesesuaiannya terhadap alur kerja dalam penelitian ini, Jupyter Notebook tidak hanya memberikan efisiensi teknis, tetapi juga mendukung kualitas dokumentasi ilmiah yang baik dan transparan [72].