BAB II
LANDASAN TEORI

2.1 Penelitian Terdahulu

Penelitian terdahulu dalam penelitian ini dikumpulkan melalui pencarian dengan beberapa kata kunci, seperti analisis sentimen, opini publik, kelapa sawit, isu lingkungan, serta Support Vector Machine, Naïve Bayes, dengan rentang tahun publikasi dari 2023 hingga 2026. Hasil dari penelitian-penelitian yang relevan tersebut kemudian dirangkum dan dijadikan sebagai referensi, yang disajikan dalam bentuk Tabel 2.1.

Tabel 2.1 Penelitian Terdahulu Analisis sentimen terhadap isu lingkungan di Indonesia pada media sosial Twitter

Tahun	Penelitian	Masalah Penelitian	Metode yang Digunakan	Hasil Penelitian
2024	Nugroho & Hasan. "Analisis Sentimen Kegiatan Pembersihan Sampah Pada Media Sosial X Menggunakan SVM dan Naïve Bayes" [24]	Menganalisis sentimen masyarakat terhadap kegiatan pembersihan sampah yang dilakukan Pandawara Group berdasarkan data dari media sosial X.	Naive Bayes dan SVM	Penelitian menggunakan 2.339 tweet dan menunjukkan bahwa SVM memperoleh akurasi 91,67%, sedangkan Naïve Bayes memperoleh akurasi 63,89%, sehingga SVM lebih efektif dalam mengklasifikasikan sentimen pada isu lingkungan dibandingkan Naïve Bayes.
2025	Rahmadani et al. "Deteksi Sentimen Publik terhadap Isu Lingkungan di Platform X (Twitter) Menggunakan Naïve Bayes dan Support Vector Machine untuk Mendukung SDGs 13 : Climate Action Public Sentiment Detection on Environmental Issues	Menganalisis sentimen publik terhadap isu lingkungan di media sosial Twitter untuk mendukung SDGs 13 (Climate Action)	Support Vector Machine (SVM) dan Naive Bayes	Metode Naive Bayes mampu mengklasifikasikan sentimen opini publik dengan cukup baik berdasarkan data Twitter.

	on Platform X (Twitter) Using Naive" [17]			
2025	Pusvita et al. "Implementasi Machine Learning untuk Analisis Sentimen Opini Publik Mengenai Pembukaan Kebun Kelapa Sawit di Papua" [14]	Menganalisis opini publik terkait pembukaan kebun kelapa sawit di Papua menggunakan pendekatan machine learning	Machine Learning SVM (Multilingual Sentiment dan Emotion Model)	Sentimen netral mendominasi sebesar 39,78%, diikuti sentimen negatif sebesar 32,16% dan positif sebesar 28,05%.
2023	Qi & Shabrina "Sentiment analysis using Twitter data: a comparative application of lexicon- and machine-learning-based approach" [3]	Membandingkan metode lexicon-based dan machine learning dalam analisis sentimen menggunakan data Twitter	Machine Learning dan Lexicon-based	Metode machine learning terbukti lebih efektif dibandingkan pendekatan lexicon-based dalam analisis sentimen data Twitter.
2025	Ardiansyah et al. "Multi-Label Classification for Opinion Mining in The Presidential Election using TF-IDF with NB And SVM" [25]	Menganalisis opini publik dalam pemilihan presiden dengan pendekatan multi-label untuk mengklasifikasikan sentimen dan kandidat secara bersamaan	Naïve Bayes dan Support Vector Machine (SVM) dengan TF-IDF	Metode SVM memiliki akurasi sekitar 93–94%, lebih tinggi dibandingkan Naive Bayes yang berada pada kisaran 86–89%.
2025	Rozia et al. "Comparison of Feature Extraction in Support Vector Machine (SVM) Based Sentiment Analysis System" [26]	Menganalisis performa metode SVM dalam klasifikasi sentimen dengan membandingkan teknik ekstraksi fitur seperti TF-IDF, Bag of Words, dan Word2Vec	Support Vector Machine (SVM) dengan TF-IDF, BoW, dan Word2Vec	Penggunaan TF-IDF pada metode SVM menghasilkan akurasi sekitar 79%, lebih tinggi dibandingkan BoW dan Word2Vec.
2025	Anam et al. "Sentiment Analysis Optimization Using Ensemble of Multiple SVM Kernel Functions" [27]	Mengoptimalkan performa analisis sentimen pada data Twitter yang tidak seimbang dengan meningkatkan akurasi klasifikasi	Support Vector Machine (SVM) dengan ensemble multi-kernel (Linear, RBF, Polynomial, Sigmoid), TF-IDF, dan SMOTE	Model SVM dengan pendekatan ensemble multi-kernel mampu mencapai akurasi hingga 98%.
2025	Sholekhah & Muntahanah. *Comparison of Naïve Bayes and Support*	Menganalisis sentimen masyarakat terhadap isu penyitaan aset koruptor	*Naïve Bayes, Support Vector Machine (SVM), TF-IDF, SMOTE*	Penelitian menggunakan 1.561 tweet dan menunjukkan bahwa SVM memperoleh akurasi

	Vector Machine in Sentiment Analysis of Confiscation of Corrupt Assets on Twitter [28]	berdasarkan data Twitter/X.		70,51% dan F1-score 0,69, sedangkan *Naïve Bayes* memperoleh akurasi 66,34% dan F1-score 0,66, sehingga SVM memiliki performa yang lebih baik dibandingkan *Naïve Bayes*.
2024	Firdaus & Suryono. *Perbandingan Algoritma Naïve Bayes dan SVM dalam Analisis Sentimen Pengguna AI di Platform X* [21]	Menganalisis sentimen pengguna terhadap teknologi *Artificial Intelligence* (AI) pada Platform X dan membandingkan performa algoritma *Naïve Bayes* dan *Support Vector Machine*.	*Naïve Bayes*, *Support Vector Machine* (SVM), TF-IDF, SMOTE	Penelitian menggunakan 9.183 tweet dan menunjukkan bahwa SVM memperoleh akurasi 96%, sedangkan *Naïve Bayes* memperoleh akurasi 84%, sehingga SVM memberikan performa klasifikasi yang lebih baik dibandingkan *Naïve Bayes*.
2025	Maulana & Darwis. *Perbandingan Metode SVM dan Naïve Bayes untuk Analisis Sentimen pada Twitter tentang Obesitas di Kalangan Gen Z* [23]	Menganalisis sentimen masyarakat terhadap isu obesitas di kalangan Generasi Z berdasarkan data Twitter serta membandingkan performa algoritma SVM dan Naïve Bayes.	Support Vector Machine (SVM), Naïve Bayes, TF-IDF	Penelitian menggunakan 4.056 tweet dan menunjukkan bahwa SVM memperoleh akurasi 89,23%, sedangkan Naïve Bayes memperoleh akurasi 72,14%, sehingga SVM memberikan performa klasifikasi yang lebih baik dibandingkan Naïve Bayes.

Tabel 2.1 menunjukkan bahwa penelitian analisis sentimen telah banyak dilakukan menggunakan berbagai pendekatan, mulai dari metode *lexicon-based*, algoritma *machine learning*, hingga model berbasis *Transformer*. Sebagian besar penelitian juga memanfaatkan data dari media sosial Twitter/X karena platform tersebut menyediakan opini publik yang bersifat terbuka dan *real-time*, sehingga banyak digunakan sebagai sumber data dalam analisis sentimen.

Nugroho dan Hasan [24] serta Rahmadani *et al.* [17] sama-sama membandingkan algoritma Support Vector Machine (SVM) dan Naïve Bayes pada analisis sentimen isu lingkungan. Kedua penelitian tersebut menunjukkan bahwa SVM memberikan performa yang lebih baik dibandingkan Naïve Bayes dalam mengklasifikasikan sentimen. Hasil tersebut menunjukkan bahwa SVM memiliki

kemampuan yang lebih baik dalam memisahkan data teks yang kompleks sehingga menghasilkan akurasi yang lebih tinggi dibandingkan Naïve Bayes.

Selain pemilihan algoritma, beberapa penelitian juga menunjukkan bahwa teknik ekstraksi fitur dan optimasi model berpengaruh terhadap performa klasifikasi. Rozia *et al.* [26] menunjukkan bahwa penggunaan TF-IDF menghasilkan performa yang lebih baik dibandingkan Bag of Words maupun Word2Vec pada metode SVM. Selanjutnya, Anam *et al.* [27] menunjukkan bahwa optimasi menggunakan *ensemble multi-kernel* dan SMOTE mampu meningkatkan akurasi hingga mencapai 98%. Hasil tersebut menunjukkan bahwa performa model tidak hanya dipengaruhi oleh algoritma klasifikasi, tetapi juga oleh teknik representasi fitur dan strategi optimasi yang digunakan.

Perkembangan penelitian terbaru juga mulai mengarah pada penggunaan model berbasis Transformer. Pusvita *et al.* [14] memanfaatkan *Multilingual Sentiment Model* dan *Ekman Emotion Model* untuk menganalisis opini publik mengenai pembukaan kebun kelapa sawit di Papua. Hasil penelitian menunjukkan bahwa model tersebut mampu melakukan analisis sentimen dengan baik, namun masih memiliki keterbatasan dalam memahami sarkasme, ironi, serta konteks bahasa Indonesia. Sementara itu, Qi dan Shabrina [3] menunjukkan bahwa pendekatan *machine learning* secara umum memberikan performa yang lebih baik dibandingkan pendekatan *lexicon-based*.

Meskipun demikian, Naïve Bayes masih banyak digunakan karena memiliki proses komputasi yang sederhana, waktu pelatihan yang relatif cepat, serta mampu memberikan hasil yang baik pada data teks berdimensi tinggi. Hal tersebut terlihat pada penelitian Ardiansyah *et al.* [25], Sholekhah dan Muntahanah [28], Firdaus dan Suryono [21], serta Maulana dan Darwis [23] yang tetap menggunakan Naïve Bayes sebagai salah satu algoritma pembanding. Namun, seluruh penelitian tersebut menunjukkan bahwa performa Naïve Bayes masih berada di bawah SVM, baik dari sisi akurasi maupun F1-score. Kondisi tersebut menunjukkan bahwa SVM secara konsisten memberikan performa klasifikasi yang lebih baik, sedangkan Naïve

Bayes tetap relevan sebagai algoritma pembanding karena memiliki kompleksitas komputasi yang lebih rendah.

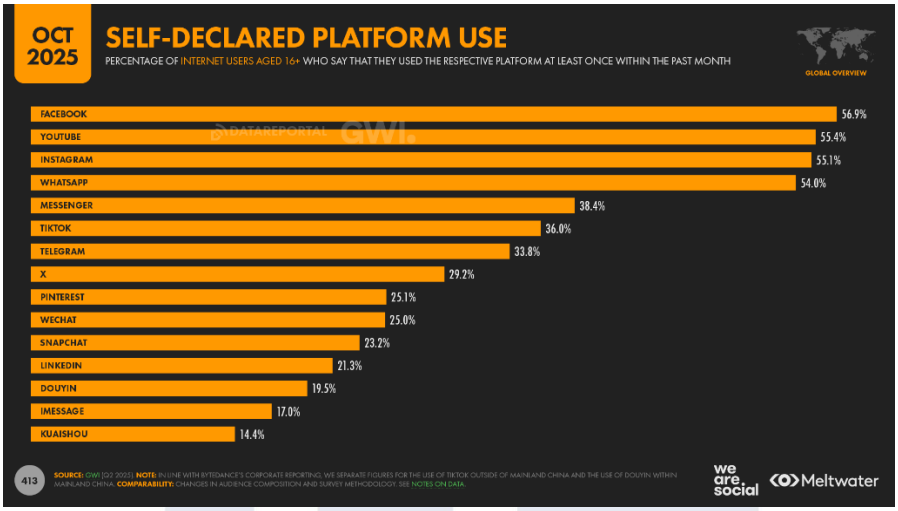
Berdasarkan penelitian-penelitian terdahulu, masih terdapat dua celah penelitian yang menjadi dasar penelitian ini. Pertama, sebagian besar penelitian analisis sentimen menggunakan domain politik, teknologi, kesehatan, maupun isu lingkungan secara umum, sedangkan penelitian yang secara khusus membahas opini publik mengenai isu lingkungan kelapa sawit di Indonesia masih sangat terbatas. Kedua, meskipun berbagai penelitian menunjukkan bahwa SVM umumnya menghasilkan performa yang lebih tinggi dibandingkan Naïve Bayes, belum banyak penelitian yang mengevaluasi kedua algoritma tersebut pada dataset isu lingkungan kelapa sawit dengan karakteristik bahasa Indonesia yang memiliki variasi kosakata, penggunaan bahasa tidak baku, serta distribusi kelas yang tidak seimbang. Oleh karena itu, penelitian ini memilih SVM dan Naïve Bayes untuk dibandingkan karena keduanya merupakan algoritma klasifikasi teks yang paling banyak digunakan dalam penelitian analisis sentimen serta memiliki karakteristik yang berbeda. SVM dikenal memiliki kemampuan yang baik dalam menangani data teks berdimensi tinggi dan menemukan batas pemisah antarkelas secara optimal, sedangkan Naïve Bayes memiliki proses komputasi yang sederhana dan efisien sehingga sering dijadikan sebagai algoritma pembanding. Selain membandingkan kedua algoritma tersebut, penelitian ini juga menerapkan TF-IDF sebagai teknik ekstraksi fitur, RandomOverSampler untuk mengatasi ketidakseimbangan data, serta validasi label secara manual untuk meningkatkan kualitas data sebelum proses klasifikasi.

2.2 Teori yang berkaitan

2.2.1 Opini Publik dan Media Sosial

Opini Publik adalah pandangan atau persepsi masyarakat tentang suatu isu yang berkembang di ruang publik [29]. Di era digital, media sosial dapat dijadikan sebagai ruang untuk mengungkapkan pendapat secara cepat oleh masyarakat [30]. Media sosial memudahkan pengguna untuk berbagi

informasi, pendapat, dan persepsi tentang berbagai isu, termasuk isu lingkungan [31].



Gambar 2.1 Penggunaan *platform* media sosial pada 2025 [32]

Grafik pada Gambar 2.1 menunjukkan tingkat penggunaan berbagai platform media sosial secara global pada tahun 2025 [32]. Terlihat bahwa platform seperti Facebook, YouTube, Instagram, dan WhatsApp masih mendominasi penggunaan, sementara platform X (Twitter) juga memiliki tingkat penggunaan yang cukup signifikan dibandingkan platform lainnya. Hal ini menunjukkan bahwa media sosial, termasuk Twitter, merupakan ruang yang aktif digunakan oleh masyarakat dalam berinteraksi dan menyampaikan opini.

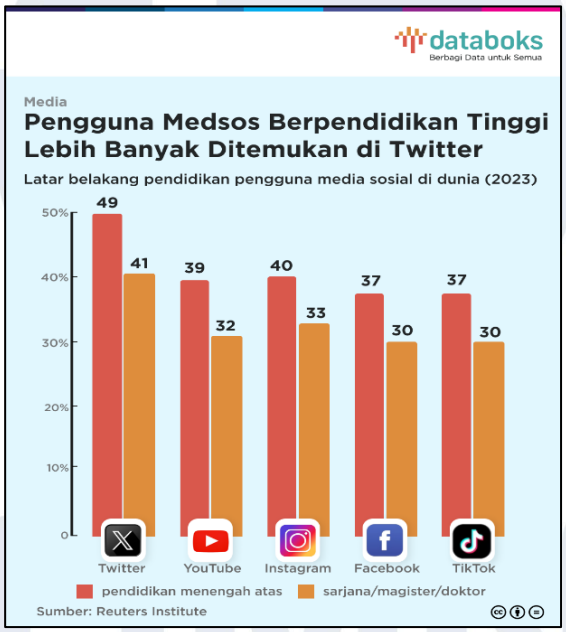
Meskipun persentase pengguna X tidak sebesar beberapa platform lain, karakteristiknya yang berbasis teks, terbuka, dan real-time menjadikannya sangat relevan sebagai sumber data dalam analisis opini publik. Oleh karena itu, data yang dihasilkan dari platform ini dapat digunakan untuk memahami persepsi masyarakat terhadap berbagai isu, termasuk isu lingkungan kelapa sawit.

Studi juga menunjukkan bahwa media sosial memiliki peran penting dalam membentuk kesadaran dan persepsi publik terhadap isu lingkungan global [31]. Diskursus yang berkembang di media sosial dapat memengaruhi cara masyarakat memahami dampak dari suatu aktivitas industri, termasuk industri

kelapa sawit [33]. Dengan demikian, media sosial tidak hanya berfungsi sebagai sarana komunikasi, tetapi juga sebagai sumber data yang potensial untuk menganalisis opini publik secara lebih luas dan mendalam.

2.2.2 Twitter/X sebagai Sumber Data Opini Publik

Twitter yang sekarang berubah menjadi X, adalah platform media sosial berbasis teks yang memungkinkan penggunanya untuk mengekspresikan pendapat mereka secara singkat dan real-time [34]. Karakteristik data yang terbuka serta sifat percakapan publik menjadikan Twitter sebagai salah satu sumber data yang banyak digunakan dalam penelitian analisis sentimen.



Gambar 2.2 Grafik pengguna medsos berpendidikan tinggi [35]

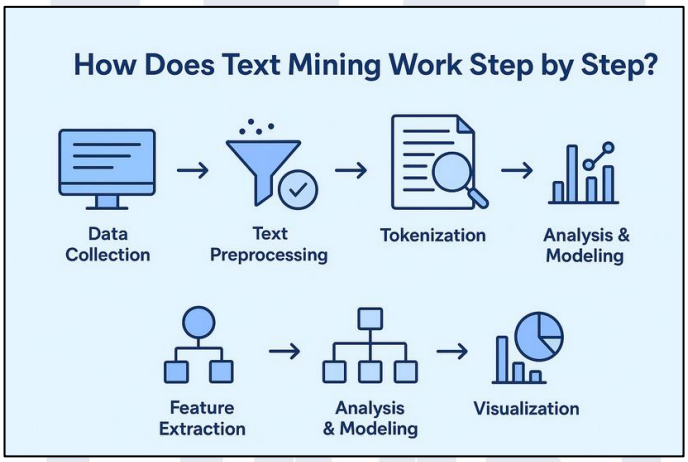
Berdasarkan Gambar 2.2 yang hasil penelitian yang dilakukan pada tahun 2024 menunjukkan bahwa, pengguna Twitter didominasi oleh kelompok dengan tingkat pendidikan yang relatif tinggi dibandingkan platform media sosial lainnya [35]. Hal ini menunjukkan bahwa opini yang disampaikan di Twitter cenderung lebih reflektif dan argumentatif, sehingga memiliki nilai yang lebih kuat dalam merepresentasikan pandangan publik terhadap suatu isu.

Studi juga menunjukkan bahwa data Twitter mampu merepresentasikan opini publik secara luas, di mana tweet yang dihasilkan mencerminkan

persepsi, sikap, serta emosi masyarakat terhadap berbagai isu publik [34], [31], [32]. Oleh karena itu, Twitter dapat digunakan sebagai sumber data yang relevan dan representatif dalam menganalisis opini publik, khususnya terkait isu lingkungan kelapa sawit.

2.2.3 Text Mining

Text mining adalah proses ekstraksi informasi dari data teks yang tidak terstruktur menggunakan teknik statistik, linguistik, dan machine learning [36]. Text mining bertujuan menemukan pola, informasi, dan pengetahuan dari kumpulan dokumen teks [37], [38]. Proses text mining meliputi beberapa tahapan, yaitu pengumpulan data, preprocessing, representasi fitur, klasifikasi, dan evaluasi.



Gambar 2.3 *Text mining steps* [39]

Gambar 2.3 menjelaskan alur kerja text mining yang dimulai dari tahap pengumpulan data teks [39], misalnya dari media sosial, lalu dilanjutkan ke preprocessing untuk membersihkan dan merapikan data supaya lebih siap diolah. Setelah itu dilakukan tokenisasi untuk memecah teks menjadi kata-kata, kemudian masuk ke tahap feature extraction agar data teks bisa diubah menjadi bentuk numerik. Selanjutnya data tersebut diproses pada tahap analisis dan modeling menggunakan algoritma machine learning seperti SVM atau Naive Bayes, dan hasil akhirnya ditampilkan dalam bentuk visualisasi agar lebih mudah dipahami. Pada data media sosial seperti Twitter, text mining

digunakan untuk mengidentifikasi pola opini publik terhadap suatu isu berdasarkan teks yang dihasilkan pengguna [40], [3].

2.2.4 Natural Language Processing (NLP)

Pemrosesan bahasa alami manusia oleh komputer adalah inti dari bidang kecerdasan buatan yang dikenal sebagai pemrosesan bahasa alami (NLP) [41]. Proses natural language processing (NLP) memungkinkan komputer untuk memahami, menganalisis, dan mengolah teks bahasa manusia [42].

Dalam analisis sentimen media sosial, NLP digunakan untuk memproses teks mentah menjadi format yang dapat dipelajari oleh algoritma pembelajaran mesin. Tokenisasi, normalisasi, penghapusan stopword, dan stemming adalah beberapa tahapan proses pengolahan bahasa natural [43].

2.2.5 Analisis Sentimen

Analisis sentimen adalah proses mengidentifikasi dan mengkategorikan opini dalam teks untuk menentukan polaritas sentimen, seperti positif, negatif, atau netral. Tujuan analisis sentimen adalah untuk memahami persepsi atau sikap penulis terhadap suatu objek atau masalah tertentu [44]. Dalam dunia media sosial, data sentimen digunakan untuk memahami sikap masyarakan terhadap suatu masalah. Penelitian menunjukkan bahwa berbagai sentimen dan emosi dapat ditemukan dalam unggahan media sosial yang berkaitan dengan lingkungan [45], [46].

2.2.6 Lexicon-Based Sentiment Analysis

Lexicon-based sentiment analysis merupakan salah satu metode dalam analisis sentimen yang digunakan untuk menentukan polaritas suatu teks berdasarkan kamus kata (*lexicon*) yang telah memiliki nilai sentimen tertentu [47],[48]. Metode ini bekerja dengan cara mencocokkan kata-kata yang terdapat dalam teks dengan daftar kata yang telah dikategorikan sebagai kata positif atau negatif, kemudian menghitung skor sentimen untuk menentukan kecenderungan opini dari teks tersebut [49].

Dalam prosesnya, setiap kata dalam teks akan diberikan bobot atau skor sesuai dengan nilai yang terdapat dalam kamus sentiment [50]. Kata yang termasuk dalam kategori positif akan memberikan kontribusi nilai positif, sedangkan kata yang termasuk dalam kategori negatif akan memberikan nilai negatif [51]. Selanjutnya, seluruh skor dari kata-kata dalam satu teks dijumlahkan untuk menentukan label sentimen akhir, apakah termasuk positif, negatif, atau netral [52], [53].

Metode lexicon-based memiliki keunggulan karena tidak memerlukan data latih dalam jumlah besar dan relatif mudah untuk diimplementasikan [54]. Namun, metode ini juga memiliki keterbatasan, seperti kurang mampu menangkap konteks kalimat secara mendalam, terutama pada penggunaan bahasa yang bersifat ambigu, ironi, atau sarkasme [55]. Dalam penelitian ini, pendekatan lexicon-based digunakan pada tahap pelabelan data untuk memberikan label sentimen awal sebelum dilakukan proses klasifikasi menggunakan algoritma machine learning [56].

2.2.7 Grid Search

Grid Search merupakan metode yang digunakan untuk mencari kombinasi pengaturan model yang paling sesuai agar dapat menghasilkan performa klasifikasi yang lebih baik[57]. Metode ini bekerja dengan mencoba berbagai kombinasi nilai yang telah ditentukan sebelumnya, kemudian membandingkan hasil yang diperoleh dari setiap kombinasi tersebut. Melalui proses ini, model dapat menemukan pengaturan yang paling optimal sehingga mampu meningkatkan kualitas prediksi dibandingkan jika hanya menggunakan pengaturan awal yang tersedia secara otomatis[58].

Pada penelitian ini, Grid Search digunakan untuk meningkatkan performa algoritma Support Vector Machine (SVM) dan Naïve Bayes dalam mengklasifikasikan sentimen. Proses pencarian dilakukan dengan menguji beberapa kombinasi pengaturan pada pembobotan TF-IDF maupun algoritma klasifikasi, kemudian mengevaluasi hasilnya menggunakan beberapa pembagian data latih secara bergantian untuk memperoleh hasil yang lebih

konsisten. Kombinasi yang menghasilkan nilai evaluasi terbaik selanjutnya digunakan untuk membangun model akhir. Dengan adanya proses ini, diharapkan model dapat mengklasifikasikan sentimen opini publik terhadap isu lingkungan kelapa sawit di Indonesia dengan tingkat akurasi yang lebih baik.

2.2.8 SMOTE

SMOTE (*Synthetic Minority Over-sampling Technique*) merupakan salah satu teknik yang digunakan untuk menangani ketidakseimbangan data (*imbalanced data*) pada proses klasifikasi [59]. Ketidakseimbangan data terjadi ketika jumlah data pada suatu kelas lebih sedikit dibandingkan kelas lainnya. Kondisi tersebut dapat membuat model lebih cenderung mengenali kelas mayoritas, sedangkan kelas minoritas menjadi kurang terklasifikasi dengan baik [60].

SMOTE bekerja dengan cara membentuk data sintetis baru pada kelas minoritas berdasarkan pola dari data yang sudah ada. Data sintetis tersebut dibuat dengan mempertimbangkan kedekatan antar data dalam kelas minoritas, sehingga jumlah data pada kelas tersebut menjadi lebih seimbang [61]. Dalam penelitian ini, SMOTE digunakan untuk membantu meningkatkan kemampuan model dalam mengenali kelas sentimen yang jumlah datanya lebih sedikit, sehingga proses klasifikasi

2.2.9 RandomOverSampler

RandomOverSampler merupakan teknik *oversampling* yang digunakan untuk mengatasi ketidakseimbangan jumlah data antar kelas [62], [47]. Teknik ini bekerja dengan cara menggandakan data dari kelas minoritas secara acak hingga jumlahnya menjadi lebih seimbang dengan kelas mayoritas [63]. Berbeda dengan SMOTE yang membentuk data sintetis baru, RandomOverSampler hanya menduplikasi data yang sudah ada pada kelas minoritas [59], [64].

Dalam penelitian ini, RandomOverSampler digunakan sebagai salah satu teknik penyeimbangan data pada proses klasifikasi sentimen. Penggunaan teknik ini bertujuan agar model tidak terlalu dominan mempelajari kelas

dengan jumlah data yang lebih banyak. Dengan jumlah data yang lebih seimbang, model diharapkan dapat mengenali pola dari masing-masing kelas sentimen, baik positif, negatif, maupun netral, secara lebih baik.

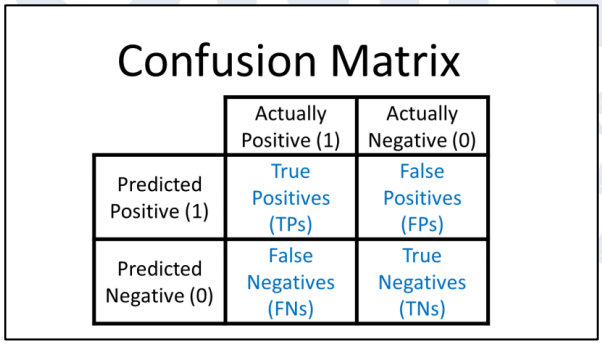
2.2.10 TF-IDF

TF-IDF merupakan metode yang digunakan untuk mengubah data teks menjadi bentuk numerik agar dapat diproses oleh algoritma machine learning [65]. Metode ini digunakan untuk mengukur seberapa penting suatu kata dalam sebuah dokumen dibandingkan dengan keseluruhan dokumen.

TF-IDF terdiri dari dua komponen utama, yaitu Term Frequency (TF) dan Inverse Document Frequency (IDF) [66]. TF digunakan untuk menghitung seberapa sering suatu kata muncul dalam sebuah dokumen, sedangkan IDF digunakan untuk mengukur seberapa unik atau jarang kata tersebut muncul di seluruh dokumen [67].

2.2.11 Evaluasi Kinerja Klasifikasi

Evaluasi model klasifikasi dilakukan untuk mengetahui seberapa baik algoritma dalam mengklasifikasikan data secara benar. Tahap ini penting karena hasil dari model tidak hanya dilihat dari prosesnya, tetapi juga dari seberapa akurat dan konsisten model dalam memberikan prediksi [68]. Salah satu metode evaluasi yang paling umum digunakan adalah confusion matrix, yaitu tabel yang membandingkan hasil prediksi model dengan kondisi sebenarnya.



Gambar 2.4 *Confusion Matrix* [69]

Gambar 2.4 menunjukkan bahwa confusion matrix terdiri dari empat komponen utama, yaitu True Positive (TP), True Negative (TN), False Positive (FP), dan False Negative (FN) [69]. TP menunjukkan data yang diprediksi benar sebagai kelas positif, TN adalah data yang benar sebagai kelas negatif, FP adalah data yang sebenarnya negatif tetapi diprediksi positif, dan FN adalah data yang sebenarnya positif tetapi diprediksi negatif.

Berdasarkan confusion matrix tersebut, terdapat beberapa metrik yang digunakan untuk mengukur performa model, yaitu *accuracy, precision, recall*, dan *F1-score*. Rumus masing-masing metrik ditunjukkan pada Persamaan (1)-(4). *Accuracy* digunakan untuk melihat proporsi prediksi yang benar dari keseluruhan data, dengan rumus (1):

$$Accuracy = \frac{(TP+TN)}{(TP+FP+FN+TN)} \quad (1)$$

Rumus 2.1 *Accuracy*

Precision digunakan untuk mengukur ketepatan prediksi pada suatu kelas, dengan rumus (2):

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

Rumus 2.2 *Precision*

Recall digunakan untuk mengukur kemampuan model dalam menemukan seluruh data pada kelas tertentu, dengan rumus (3):

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

Rumus 2.3 *Recall*

Sedangkan *F1-score* merupakan rata-rata harmonik antara precision dan recall, yang dirumuskan sebagai (4):

$$F1 = 2 \times \frac{Precision \times Recall}{Precision+Recall} \quad (4)$$

Rumus 2.4 *F1-Score*

Dengan adanya metrik evaluasi tersebut, performa model seperti SVM atau Naïve Bayes dapat dianalisis secara lebih menyeluruh, sehingga dapat diketahui apakah model sudah cukup baik dalam mengklasifikasikan sentimen opini publik pada data Twitter atau masih perlu dilakukan perbaikan.

2.2.12 Pembersihan Data / *Data Cleaning*

Data cleaning atau pembersihan data merupakan salah satu tahapan penting dalam proses pengolahan data yang bertujuan untuk meningkatkan kualitas data sebelum dilakukan analisis [70]. Data yang diperoleh dari berbagai sumber, terutama media sosial, umumnya masih mengandung informasi yang tidak relevan, data duplikat, data kosong (*missing value*), maupun karakter yang tidak diperlukan. Apabila data tersebut tidak dibersihkan, hasil analisis maupun proses pembentukan model dapat menjadi kurang optimal [71].

Proses *data cleaning* dilakukan untuk mengidentifikasi dan memperbaiki kualitas data sehingga data yang digunakan menjadi lebih konsisten, akurat, dan sesuai dengan kebutuhan penelitian. Tahapan ini dapat meliputi penghapusan data duplikat, penanganan *missing value*, penghapusan data yang tidak relevan, serta perbaikan format data apabila diperlukan. Dengan melakukan *data cleaning*, data yang akan digunakan pada tahap selanjutnya menjadi lebih terstruktur dan dapat mengurangi kemungkinan terjadinya kesalahan selama proses analisis maupun klasifikasi.

Dalam penelitian yang menggunakan data teks dari media sosial, *data cleaning* juga berperan dalam menghilangkan informasi yang tidak memberikan kontribusi terhadap proses analisis, sehingga kualitas data menjadi lebih baik dan representasi informasi yang dihasilkan menjadi lebih akurat.

2.2.13 Text Preprocessing

Text preprocessing merupakan tahapan awal dalam *text mining* yang bertujuan untuk mengubah data teks mentah menjadi data yang lebih bersih, terstruktur, dan siap digunakan pada proses analisis maupun klasifikasi [72].

Data teks yang berasal dari media sosial umumnya masih mengandung berbagai unsur yang tidak diperlukan, seperti penggunaan huruf besar dan kecil yang tidak konsisten, tanda baca, simbol, angka, kata tidak baku, maupun kata-kata yang tidak memiliki makna penting dalam proses klasifikasi [73]. Oleh karena itu, diperlukan serangkaian tahapan *preprocessing* agar kualitas data dapat ditingkatkan.

Secara umum, *text preprocessing* terdiri atas beberapa tahapan, yaitu *case folding*, *tokenizing*, normalisasi kata, *stopword removal*, dan *stemming* [74]. *Case folding* dilakukan untuk menyeragamkan seluruh huruf menjadi huruf kecil sehingga tidak terjadi perbedaan penulisan kata yang memiliki makna sama. Selanjutnya, *tokenizing* digunakan untuk memisahkan kalimat menjadi kumpulan kata atau *token*. Tahap normalisasi bertujuan mengubah kata tidak baku, singkatan, maupun kata yang mengalami variasi penulisan menjadi bentuk yang baku.

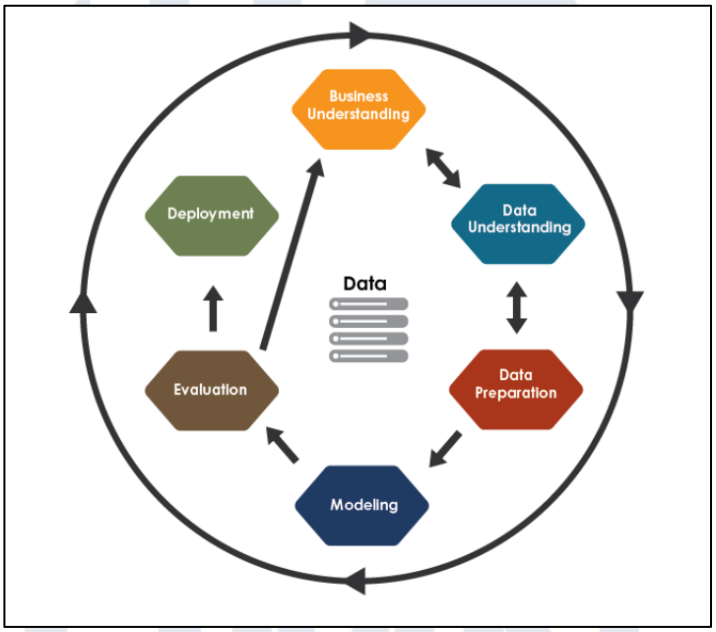
Tahap berikutnya adalah *stopword removal*, yaitu menghapus kata-kata yang sering muncul tetapi tidak memberikan informasi penting terhadap proses analisis, seperti kata hubung atau kata depan. Setelah itu, dilakukan *stemming* untuk mengubah setiap kata menjadi bentuk dasarnya sehingga kata-kata yang memiliki imbuhan dapat direpresentasikan dalam bentuk yang sama. Melalui serangkaian tahapan tersebut, data teks menjadi lebih konsisten dan mampu menghasilkan representasi fitur yang lebih baik pada proses ekstraksi fitur maupun klasifikasi.

Penerapan *text preprocessing* tidak hanya bertujuan untuk mengurangi *noise* pada data, tetapi juga membantu meningkatkan kualitas representasi data sehingga algoritma *machine learning* dapat mempelajari pola data secara lebih efektif. Dengan demikian, proses klasifikasi yang dilakukan diharapkan mampu menghasilkan performa yang lebih baik dibandingkan apabila menggunakan data teks yang belum melalui tahap *preprocessing*.

2.3 *Framework/Algoritma* yang digunakan

2.3.1 *Cross-Industry Standard Process for Data Mining* (CRISP-DM)

Cross-Industry Standard Process for Data Mining (CRISP-DM) merupakan salah satu *framework* yang banyak digunakan dalam penelitian berbasis *data mining* karena menyediakan tahapan yang sistematis dan terstruktur dalam proses pengolahan data hingga menghasilkan sebuah model [75]. *Framework* ini dikembangkan oleh *European Strategic Programme on Research in Information Technology* (ESPRIT) pada tahun 1996 [76] dan hingga saat ini masih banyak diterapkan pada berbagai penelitian maupun pengembangan sistem yang memanfaatkan *data mining* dan *machine learning* [77].



Gambar 2.5 Proses CRISP-DM [78]

Gambar 2.5 adalah proses CRISP-DM. CRISP-DM terdiri dari enam tahapan utama, yaitu *Business Understanding*, *Data Understanding*, *Data Preparation*, *Modeling*, *Evaluation*, dan *Deployment* [78]. Keenam tahapan tersebut saling berkaitan dan dapat dilakukan secara berulang apabila pada proses evaluasi masih ditemukan hasil yang belum sesuai dengan tujuan penelitian. Dengan adanya alur kerja yang terstruktur, CRISP-DM membantu peneliti dalam mengelola data secara lebih sistematis sehingga proses pengembangan model menjadi lebih terarah.

Tahap pertama, yaitu *Business Understanding*, bertujuan untuk memahami permasalahan yang akan diselesaikan, menentukan tujuan penelitian, serta merumuskan kebutuhan analisis. Tahap kedua, *Data Understanding*, dilakukan untuk mengumpulkan, mempelajari, dan memahami karakteristik data yang akan digunakan, termasuk mengidentifikasi kualitas data maupun permasalahan yang terdapat pada *dataset*.

Tahap ketiga adalah *Data Preparation*, yaitu proses menyiapkan data agar siap digunakan dalam proses pemodelan. Tahap ini umumnya meliputi pembersihan data, transformasi data, seleksi atribut, serta proses *preprocessing* lainnya sesuai kebutuhan penelitian. Setelah data siap digunakan, proses dilanjutkan ke tahap *Modeling*, yaitu membangun model menggunakan algoritma yang dipilih serta melakukan penyesuaian parameter apabila diperlukan.

Tahap kelima adalah *Evaluation*, yaitu mengevaluasi performa model untuk mengetahui apakah hasil yang diperoleh telah memenuhi tujuan penelitian. Evaluasi dilakukan menggunakan metrik yang sesuai dengan jenis permasalahan, sehingga dapat diketahui kelebihan maupun kekurangan dari model yang dibangun. Tahap terakhir adalah *Deployment*, yaitu penerapan hasil model ke dalam suatu sistem atau lingkungan penggunaan sehingga hasil analisis dapat dimanfaatkan dalam pengambilan keputusan maupun kebutuhan lainnya.

Melalui tahapan-tahapan tersebut, CRISP-DM menjadi salah satu *framework* yang banyak digunakan karena bersifat fleksibel [79], mudah diterapkan pada berbagai jenis data, serta mampu memberikan alur kerja yang jelas dalam proses pengembangan model *data mining* dan *machine learning*.

2.3.2 SVM (Support Vector Machine)

Support Vector Machine (SVM) merupakan salah satu algoritma machine learning yang banyak digunakan dalam berbagai bidang, termasuk lingkungan dan kesehatan, karena kemampuannya dalam melakukan klasifikasi data secara efektif [80], [81]. Dalam proses klasifikasi, SVM bekerja dengan membentuk

sebuah hyperplane optimal yang mampu memisahkan data ke dalam kelas-kelas yang berbeda. Untuk menentukan kelas dari suatu data baru, SVM menggunakan fungsi keputusan (decision function) yang dirumuskan secara matematis seperti pada Rumus 2.1.

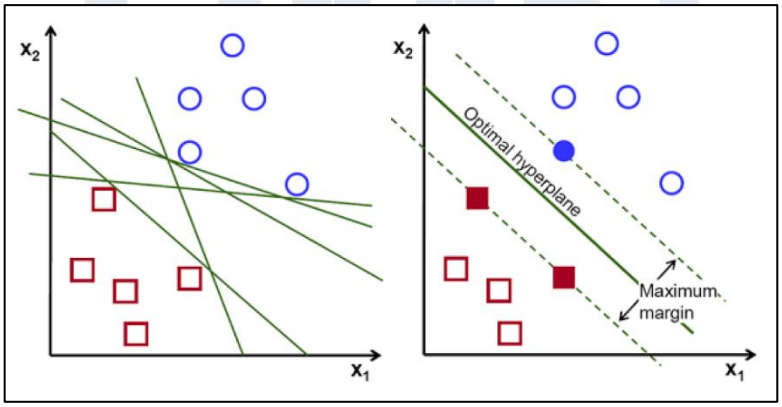
$$f(x) = sign(\sum_{i=1}^{n} y_i \alpha_i K(x_i, x) + b)$$

Rumus 2.5 Rumus *Support Vector Machine*

Penjelasan singkat dari setiap komponen Adalah sebagai berikut.

- x = data yang mau diprediksi
- x_i = support vector (data penting dari training)
- y_i = label kelas (-1 atau +1)
- α_i = bobot hasil training
- $K(x_i, x)$ = fungsi kernel (misalnya linear, RBF, dll)
- b = bias
- **sign()** = menentukan kelas akhir

Secara konseptual, fungsi keputusan pada Rumus 2.1 tersebut merepresentasikan proses penentuan posisi suatu data terhadap *hyperplane* pemisah. Visualisasi dari konsep ini dapat dilihat pada Gambar 2.6.



Gambar 2.6 *Hyperplane SVM* [82]

Gambar 2.6 terlihat bahwa sebenarnya ada banyak garis yang bisa memisahkan dua kelas [82]. Namun, SVM memilih garis terbaik yang

memiliki jarak paling besar terhadap titik terdekat dari masing-masing kelas (maximum margin). Titik yang paling dekat dengan garis tersebut disebut *support vector*, dan titik inilah yang menentukan posisi garis pemisah.

2.3.3 Naïve Bayes

Naïve Bayes merupakan salah satu algoritma klasifikasi yang berbasis pada Teorema Bayes dan sering digunakan dalam analisis sentimen karena sederhana serta cukup efektif dalam mengolah data teks [83]. Algoritma ini bekerja dengan menghitung probabilitas suatu data termasuk ke dalam kelas tertentu berdasarkan fitur yang dimiliki [84]. Secara matematis, *Naïve Bayes* dinyatakan dengan persamaan pada rumus 2.2.

$$P(c|x) = \frac{P(x|c)\cdot P(c)}{P(x)}$$

Rumus 2.6 Rumus *Naïve Bayes*

- x = Data dengan class yang belum diketahui
- c = Hipotesis data merupakan suatu class spesifik
- $P(c|x)$ = probabilitas data $x$termasuk ke kelas c (*posterior*)
- $P(x \mid c)$ = probabilitas data $x$jika diketahui kelasnya c (*likelihood*)
- $P(c)$ = probabilitas awal suatu kelas (*prior*)
- $P(x)$ = probabilitas kemunculan data x (*evidence*)

Rumus 2.2 menampilkan $P(c \mid x)$, yang merupakan probabilitas suatu data termasuk ke dalam kelas tertentu, $P(c)$adalah probabilitas awal dari kelas, $P(x \mid c)$adalah *likelihood* atau probabilitas kemunculan data pada kelas tersebut, dan $P(x)$adalah probabilitas dari data [85].

Naïve Bayes menganggap setiap fitur itu berdiri sendiri terhadap kelas, jadi antar fitur tidak saling mempengaruhi. Walaupun asumsi ini cukup sederhana, metode ini tetap bisa memberikan hasil yang baik, terutama untuk data teks seperti analisis sentimen. Dibandingkan dengan *ZeroR* yang tidak memakai fitur atau *OneR* yang hanya pakai satu fitur, *Naïve Bayes* memanfaatkan semua fitur dengan pendekatan probabilitas berdasarkan *Teorema Bayes*.

2.4 Tools/software yang digunakan

2.4.1 *Phyton*

Python adalah bahasa pemrograman serbaguna yang bersifat interpretatif, dikembangkan oleh Guido van Rossum pada tahun 1991 [86]. Sebagai bahasa pemrograman gratis, Python dapat digunakan untuk berbagai keperluan komersial. Dengan sintaksis yang sederhana namun powerful, serta dukungan library yang lengkap dan fungsional, Python menjadi salah satu pilihan utama dalam pengembangan *machine learning* dan *deep learning* [87]. Kemudahan belajar serta fleksibilitasnya sebagai bahasa pemrograman tingkat tinggi menjadikan *Python* semakin populer. Menurut penggunanya, keberadaan berbagai *library* pendukung mempermudah pengguna dalam menerapkan machine learning secara lebih efisien.

2.4.2 Jupyter Notebook

Jupyter Notebook adalah aplikasi *web* interaktif yang digunakan untuk membuat dan berbagi dokumen komputasi [88]. Aplikasi ini pertama kali dikenal sebagai *IPython Notebook* sebelum berganti nama menjadi Jupyter pada tahun 2014 [89]. Notebook disimpan dalam format **.ipynb** dan memungkinkan pengguna untuk menulis serta menjalankan kode, mencatat hasilnya, dan menambahkan dokumentasi dengan markdown.

Jupyter Notebook mendukung lebih dari 40 bahasa pemrograman, termasuk Python, R, dan Scala. Fitur utama dari Jupyter *Notebook* termasuk *dashboard* untuk mengelola dokumen, kernel untuk menjalankan kode interaktif, serta kemampuan untuk mengonversi *notebook* ke berbagai format seperti HTML, LaTeX, PDF, *Markdown*, atau Python [90].

2.4.3 Microsoft Excel

Microsoft Excel merupakan salah satu tools yang digunakan untuk membantu proses pengolahan data karena memiliki fitur yang mudah digunakan dan cukup lengkap. Excel digunakan untuk memasukkan data, menyusun tabel, memeriksa data yang kosong atau duplikat, serta memperbaiki

format angka agar data lebih rapi dan siap digunakan [91]. Selain itu, Excel juga mendukung berbagai rumus seperti SUM, IF, COUNTIF, dan VLOOKUP yang dapat membantu proses perhitungan menjadi lebih cepat dan efisien.

Excel juga memiliki fitur seperti filter, sort, conditional formatting, dan pivot table yang mempermudah proses analisis data dalam jumlah besar. Melalui fitur tersebut, data dapat dirangkum menjadi informasi yang lebih mudah dipahami dan divisualisasikan dalam bentuk grafik sederhana seperti bar chart, line chart, dan pie chart. Pada kegiatan magang, Excel digunakan sebagai tools awal untuk membersihkan dan memeriksa data sebelum diolah lebih lanjut menggunakan Python dan divisualisasikan menggunakan Power BI [92].