BAB II
LANDASAN TEORI

2.1 Penelitian Terdahulu

Setelah dilakukan studi mendalam terkait dengan topik analisis sentimen, didapat studi- studi sebelumnya yang mempunyai keterkaitan dengan studi ini. Berbagai studi tersebut tentunya membahas berbagai aspek serta metodologi yang menjadi dasar dalam studi ini. Beberapa penelitian-penelitian terdahulu yang terdapat dalam tabel 2.1

Tabel 2. 1 Penelitian Terdahulu

Judul Jurnal	**Teknik Penelitian**	**Hasil Penelitian**
"Analisis Sentimen Terhadap Penggunaan Artis Korea Selatan Sebagai *Brand Ambassador* Produk Makanan Dan Minuman Menggunakan Algoritma *Support Vector Machine (SVM)*" - e- Proceeding of Engineering : Vol. 1, No.4 Agustus 2024 - Endarpariswara, Dita Pramesti Riska, Yanu Fa'rifah/ 2024[8]	SVM	Pada studi ini, pengimplementasian algoritma SVM dengan menggunakan metode SMOTE untuk penanganan imblance menghasilkan akurasi 83,89%, precision 85%, recall 80% yang menandakan bahwa algoritma tersebut mampu mengklasifikasikan sentimen dengan performa tinggi
"*Twitter sentiment analysis of South Korea artist as brand ambassadors of local beauty product*", Journal of Information Technology and Computer Science, Ristyani Slamet, Windu Gata, Annisa Novtariany, Khairunisa, Hilyati, Febri Ainun Jariyah/ 2022[9]	SVM	Pengimplementasian SVM (*Support Vector Machine*) dalam sebuah analisis sentimen, menghasilkan tingkat akurasi sebesar 83.60% yang dimana hasil ini menunjukkan bahwa model ini dapat dengan baik memprediksi sentimen positif, negatif, atau netral yang masyarakat opinikan pada *platform* Twitter.
"Analysis of Twitter user sentiment towards BTS using the Support Vector Machine" – Journal of Applied Informaticss and Computing – Tiara Safitri, Yuyun Umaidah, Iqbal Maulana/ 2023[34]	SVM	Menunjukkan bahwa algoritma Support Vector Machine berhasil mengenali dengan sentimen masyarakat terhadap grup BTS yang menjadi brand ambassador

Judul Jurnal	Teknik Penelitian	Hasil Penelitian
		Tokopedia.
"Sentiment Analysis of Indonesian Interest in Korean Food Based on Naïve Bayes Algorithm" – Jurnal Sosioteknologi, Institut Teknologi Telkom Purwokerto – Felmi Putra Pratama Subandi, Fauzan Romadlon, Isniani Nurisusilawati, Anita Chindyana/ 2022[30]	Naïve Bayes	Penggunaan metode algoritma Naïve Bayes merupakan metode yang cocok guna melakukan klasifikasi sentimen pada komentar Youtube.
"Analisis Sentimen Brand Ambassador Artis Korea Selatan Pada Produk Indonesia dengan Lexicon" – Seminar Nasional Sains Data 2023 (SENADA 2023) UPN "Veteran" Jawa Timur – Galuh Etha Pratiwi, Tiani Wahyu Utami. Rochdi Wasono/ 2023[11]	Lexicon	Penelitian ini menunjukkan bahwa penggunaan artis Korea Selatan sebagai brand ambassador produk Indonesia mendapatkan respon yang positif dari masyarakat, berdasarkan analisis sentiment menggunakan metode lexicon-based (kamus SentiWordNet). Dari total 400 komentar yang dianalisis sebagian besar termasuk dalam kategori sentiment positif, disusul oleh sentiment netral, serta sebagian kecil bersifat negatef.
"Analisis Sentimen Brand Ambassador BTS terhadap Tokopedia Menggunakan Klasifikasi Bayesian Network dengan ekstraksi fitur TF-IDF" – Jurnal Informatika Polinema – Regina, Triando Hamonangan Saragih, Dwi Kartini/ 2023[31]	Bayesian Network	Penelitian ini berhasil mengakji sentimen dengan akurasi model sebesar 87%, hal ini menunjukkan bahwa model Bayesian Network dengan TF-IDF dapat mengklasifikasikan sentimen dengan tingkat ketepatan yang tinggi.masyarakat terhadap Tokopedia yang menggandeng BTS sebagai *brand ambassador*, menggunakan metode Bayesian Network dengan fitur TF-IDF. Hasil analisis menunjukkan bahwasanya sebagian besar sentimen bersifat positif, menunjukkan dukungan publik pada kolaborasi tersebut. Model klasifikasi Bayesian Network mampu melakukan analisis sentimen secara efektif, dengan performa akurasi yang layak dalam mengklasifikasikan opini pengguna media sosial.

Judul Jurnal	Teknik Penelitian	Hasil Penelitian
		Strategi pemanfaatan BTS sebagai duta merek dinilai berdampak baik terhadap citra Tokopedia di mata konsumen, terutama di kalangan penggemar K-pop.
"Analisis Sentimen Penggemar Treasure di Karnaval Mandiri Menggunakan Naïve Bayes" – Jurnal Sistem Informasi Triguna Dharma (JURSI TGD) 3 – Chulyatunni'mah, Rudi Kurniawan, Saeful Anwar/ 2025[32]	Naïve Bayes	Studi ini menerapkan metode Naïve Bayes, menggunakan TextBlob guna pelabelan dan SMOTE untuk penyeimbangan data. Akurasi model yang di dapatkan dalam penelitian ini adalah sebesar 96% dengan 38,1% sentiment netral, 35,3% positif, 26,6% negatif.
"Analisis Sentimen Komentar Netizen Terhadap Pembubaran Konser NCT127 Menggunakan Metode Naïve Bayes" – Journal of Information Systems Research (JOSH) – Nisa Qonita Rizkina, Firman Noor Hasan/ 2023[40]	Naïve Bayes	Akurasi model mencapai 84%**,** yang menandakan bahwa metode *Naïve Bayes* cukup efektif dalam mengkelompokkan opini publik berbasis teks. Proses preprocessing dan pemodelan dilakukan secara sistematis, mencakup pengambilan data, pembersihan teks, hingga evaluasi performa model.
"Analisa Sentimen Drama Korea Melalui Media Sosial X dengan Menggunakan Algoritma Naïve Bayes – Jurnal Indonesia : Manajemen Informatika dan Komunikasi – Putri Dwi Aprilia, Sri Lestari/ 2024[41]	Naïve Bayes	Dalam penelitian ini digunakan algoritma Naïve Bayes dengan menggunakan fitur TF-IDF menghasilkan akurasi yang baik yaitu sekitar 80% keatas.
"Analisis Sentimen Keberhasilan Debut Grup K-Pop Pada Platform X Menggunakan Algoritma Support Vector Machine" – Jurnal Seminar Nasinal Corisindo – Indri Febriana Putri, Wita Witanti, Edvin Ramadhan/ 2024[36]	SVM	Pengimplementasian metode SVM dalam analisis sentimen ini menghasilkan akurasi yang sangat tinggi, penulis memilih 3 grup K-Pop untuk diteliti dan hasilnya grup K-Pop Baby Monster memiliki tingkat akurasi sebesar 95%, Zerobaseone sebesar 99%, dan grup Kiss Of Life memiliki tingkat akurasi 98%, berdasarkan dari hasil tingkat akurasi yang didapatkan, algoritma SVM sangat efektif dalam mengklasifikasikan sentiment.
"Analisis Sentimen Ulasan Aplikasi	Random Forest	Pengimplementasian algoritma

Judul Jurnal	Teknik Penelitian	Hasil Penelitian
Dana dengan Metode Random Forest" – Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer Vol. 6, No. 9, September 2022 hlm. 4305-4313 – Fanka Angeline Larasati, Dian Eka Ratnawati, Buce Trias Hanggara[37]		*Random Forest* menghasilkan algoritma akurasi 84% dengan parameter terbaik dengan kedalaman tree 65, dan jumlah tree 400. Sentimen diuji menggunakan *K-Fold Cross Validation* (5-fold)
"Penerapan Metode Random Forest Dalam Menganalisis Sentimen Pengguna Aplikasi Capcut Di Google Play Store" – Jurnal Mahasiswa Teknik Informatika Vol. 7 No. 6, Desember 2023 - Ayu Sagita, Ahmad Faqih, Gifthera Dwilestari, Bambang Siswoyo, Denni Pratama[42]	Random Forest	Studi ini menerapkan algoritma *Random Forest* untuk mengklasifikasikan sentimen pengguna terhadap aplikasi CapCut. Proses penelitian dilakukan secara terstruktur, mulai dari tahap pengumpulan data hingga evaluasi model, serta didukung oleh dasar teori yang kuat. Evaluasi performa model menerapkan *Classification Report* serta *Confusion Matrix* menunjukkan hasil akurasi sebesar 86%, presisi 89%, recall 81%, dan f1-score 85%. Capaian ini mengindikasikan bahwasannya model mempunyai kinerja yang baik dalam mendeteksi sentimen.
"An Integrated Approach for Sentiment Analysis and Topic Modeling of a Digital Bank in Indonesia using Naïve Bayes and Latent Dirichlet Allocation Algorithms on Social Media Data" – IBDAP Confrence Bangkok 2023 - Johan Setiawan, Anastasia Milenia, Ahmad Faza[38]	Naïve Bayes dan LDA	Pada penelitian ini algoritma Naïve Bayes berhasil mengklasifikasi sentiment dengan performa baik, mencapai F1- Score 0.863 terutama pada data Twitter dan LDA berhasil mengidentifikasi 4 topik utama dari sentimen positif, 5 topik utama dari sentimen negatif dan coherence score untuk topik positif sebesar 0.426 dan negatif sebesar 0.397 yang menunjukkan kualitas topik cukup baik.
"*Sentiment Analysis of Public Acceptance of Covid-19 Vaccines Types in Indonesia using Naïve Bayes, Support Vector Machine, and Long Short-Term Memory*	Naïve Bayes, SVM, LSTM	Penelitian ini membandingkan performa algoritma Naïve Bayes**,** Support Vector Machine (SVM)**,** dan Long Short-Term Memory

Judul Jurnal	Teknik Penelitian	Hasil Penelitian
(LSTM)" – Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi), vol. 7, no. 3, pp. 722-732, 2023 – Dinar Ajeng Kristiyanti, Sri Handayani[39]		(LSTM) dalam menganalisis sentimen publik terhadap berbagai jenis vaksin Covid-19 di Indonesia berdasarkan data Twitter (2.000 tweet dari Januari 2021–2022). Dengan memperoleh akurasi SVM sebesar 84,89%, Naïve Bayes 84,65%, dan LSTM sebesar 82,97%.

Tabel 2.1 merupakan penelitian terdahulu mengenai analisis sentimen terhadap penggunaan figur publik Korea Selatan sebagai *brand ambassador* telah banyak dilakukan khususnya pada media sosial seperti Twitter atau *platform* X. Dalam berbagai studi, algoritma untuk pengklasifikasian secara luas digunakan untuk mengidentifikasi dan mengelompokkan opini masyarakat menjadi sentimen positif, negatif, maupun netral.

Algoritma SVM terbukti memiliki performa tinggi dalam menangani data teks yang kompleks serta mampu menghasilkan akurasi diatas 80%[8][9][10]., khususnya ketika dikombinasikan dengan teknik *balancing* seperti SMOTE dan ekstraksi fitur TF-IDF. Meskipun algoritma SVM terbukti mampu dalam menghasilkan akurasi yang tinggi pada penelitian sebelumnya, *dataset* yang digunakan adalah dataset penggunaan artis Korea sebagai *brand ambassador* produk lokal secara spesifik. Sementara itu, algoritma Naïve Bayes banyak digunakan karena kemudahannya dalam penerapan serta efektivitasnya terhadap data komentar yang bersifat pendek dan eksplisit. Beberapa studi juga mengintegrasikan pendekatan *lexicon-based*, seperti penggunaan *SentiWordNet*, untuk menangkap sentimen berdasarkan kamus emosi yang telah ditentukan.

Penggunaan algoritma *Random Forest* sebagai bagian dari metode *ensemble* learning menunjukkan hasil akurasi yang kompetitif, dengan ketepatan klasifikasi yang cukup stabil pada data yang bervariasi. Beberapa pendekatan lanjutan juga menggabungkan metode klasifikasi dengan pemodelan topik, seperti *Latent Dirichlet Allocation* (LDA), guna memperoleh wawasan yang lebih mendalam mengenai hasil dari analisis sentimen. Secara umum, hasil-hasil penelitian tersebut

menunjukkan bahwa implementasi model klasifikasi sentimen dapat memberikan kontribusi yang signifikan dalam mengevaluasi efektivitas strategi branding, khususnya dalam memahami reaksi konsumen terhadap kolaborasi antara produk lokal dan artis Korea. Akan tetapi, sebagian besar penelitian masih berfokus pada satu jenis algoritma dan belum banyak melakukan pendekatan komparatif secara menyeluruh, serta belum menyoroti secara mendalam dinamika kontroversi dan resistensi publik terhadap strategi *branding* lintas budaya tersebut.

Oleh karena itu, penelitian ini hadir dengan pendekatan yang lebih komprehensif dengan menggunakan *dataset* yang berbeda yaitu menggunakan *dataset* penggunaan artis Korea sebagai *brand ambassador* produk lokal secara keselurahan dan tidak spesifik secara produknya dan membandingkan performa tiga algoritma yaitu *Naïve Bayes*, SVM, dan *Random Forest* dalam mengklasifikasikan sentimen masyarakat terhadap produk lokal Indonesia yang menggandeng artis Korea sebagai *brand ambassador*. Fokus utama penelitian ini bukan hanya untuk satu kategori merek produk Indonesia saja, tetapi penelitian ini menggunakan seluruh produk lokal yang menggunakan artis Korea sebagai *brand ambassador* produk lokal

2.2 Teori Penelitian

2.2.1 Brand Ambassador

Brand Ambassador merupakan seseorang ataupun sekelompok orang yang dipilih dan ditunjuk oleh suatu perusahaan yang mempunyai produk atau merek untuk menjadi duta dalam produk atau merek dari perusahaan dalam upaya mempromosikan, memasarkan, dan membangun citra positif terhadap produk ataupun merek tersebut[11]. *Brand Ambassador* biasanya merupakan seseorang yang mempunyai kepopularitasan seperti selebriti, aktor, atlet terkenal dan juga *public figure* yang mempunyai pengaruh besar dan kepopularitasan dalam suatu komunitas tertentu[12].

Seseorang atau sekelompok orang yang sudah dipilih menjadi B*rand Ambassador* terhadap suatu produk ataupun merek tentunya mempunyai tanggung jawab yang besar, mereka dituntut untuk menjaga integritas dan otentisitas dalam segala interaksi yang berhubungan dengan produk atau

merek guna agar tidak merusak citra produk ataupun merek. *Brand Ambassador* juga mempunyai banyak peran penting seperti mewakili dan menjadi wajah dari sebuah produk atau merek untuk sebuah promosi dalam media apapun. Penggunaan *Brand Ambassador* dalam sebuah produk atau merek tidak semerta-merta hanya untuk mempromosikan produk, banyak keuntungan yang dapat dihasilkan dalam penggunaan *Brand Ambassador* salah satunya adalah mempengaruhi keputusan pembelian konsumen[13]. *Brand Ambassador* pastinya merupakan seseorang yang mempunyai kepopularitasan dan mempunyai banyak penggemar, para penggemar mereka pastinya cenderung akan mencoba sebuah produk atau merek yang direkomendasikan oleh idola mereka[14].

2.2.2 X

X yang juga dikenal sebelumnya dengan Twitter, merupakan sebuah media sosial yang dijadikan sebagai ruang publik digital yang memungkinkan pengguna nya untuk melakukan diskusi, menyampaikan opini dan juga membentuk wacana publik secara terbuka[15]. *Platform* media sosial X digunakan juga sebagai media komunikasi yang memiliki kekayaan dalam menyampaikan informasi melalui teks, gambar, video, dan tautan, X memberikan fasilitas untuk penggunanya untuk pembentukan opini publik melalui penggunaan *hashtag* yang menunjukkan bahwa penyebaran informasi secara luas dan cepat. Analisis jaringan komunikasi pada X dapat menunjukkan bagaimana kelompok-kelompok pengguna dapat membentuk komunitas dan mengarahkan dikursus publik melalui interaksi dan *retweet*[15].

2.2.3 Analisis Sentimen

Analisis sentimen merupakan suatu proses guna menilai sentimen ataupun pandangan yang dinyatakan oleh seseorang melalui teks yang dapat diklasifikasikan sebagai sentimen positif, negatif, ataupun netral. *Sentiment analysis* merupakan cabang dari pemrosesan bahasa alami yang bertujuan mengidentifikasi dan mengevaluasi opini, sikap, dan emosi seseorang terhadap topik, produk, layanan, individu, atau aktivitas tertentu

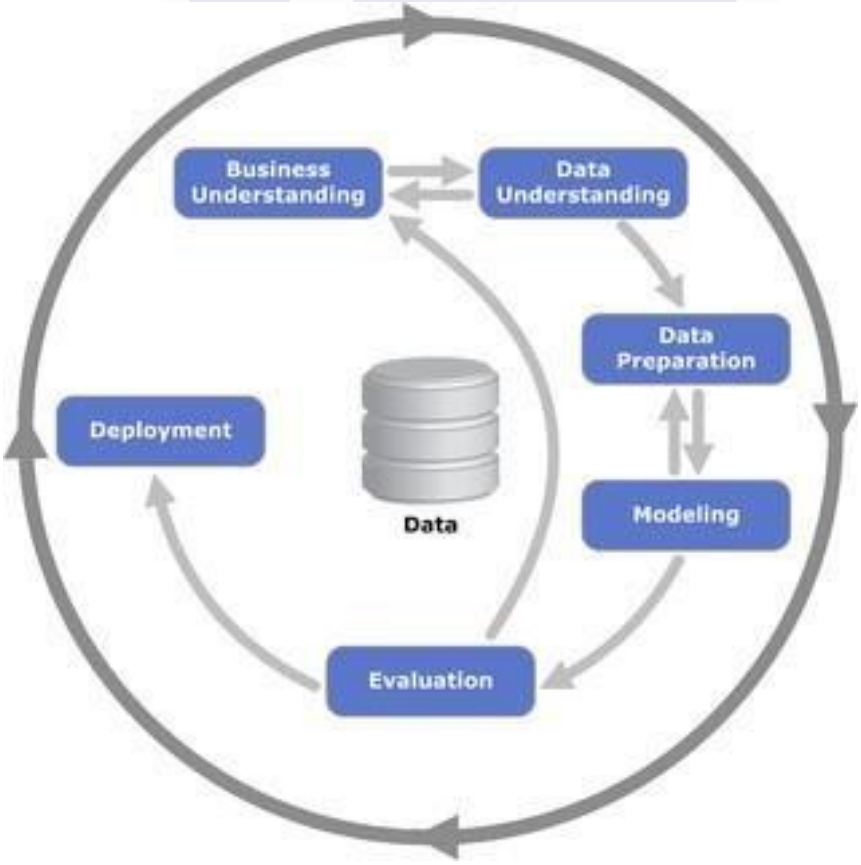
[16].

Banyaknya pengguna internet dan media sosial yang membagikan pengalaman, opini, dan perhatian mereka dalam bentuk sebuah tulisan. Tulisan yang dapat mencakup berbagai perasaan, baik positif, netral, maupun negatif, yang dapat diungkapkan dengan tingkat kompleksitas yang beragam[17].

2.3 Framework dan algoritma penelitian

2.3.1 CRISP-DM

CRISP-DM (*Cross-Industry Standard Process for Data Mining*) adalah metode standar yang digunakan dalam pengolahan data. Model ini mencakup beberapa tahapan seperti pemahaman bisnis, eksplorasi data, persiapan data, pemodelan, evaluasi model, dan penerapan hasil. Pendekatan ini banyak digunakan dalam proyek data *science* karena fleksibel dan terstruktur.[18]. Metode yang terdapat dalam CRISP-DM terdiri dari enam fase yaitu *Business Understanding*, *Data Understanding, Data Preparation, Modelling, Evaluation*, serta *Deployment*. Berikut merupakan gambar dibawah merupakan tahapan dari CRISP-DM :



Gambar 2. 1Alur Metode CRISP-DM

a. Business Understanding

Tahap *Business Understanding* ialah proses untuk memahami kebutuhan serta sasaran bisnis, Lalu dikonversi menjadi identifikasi permasalahan pada *Data Mining*. Kemudian, disusun perencanaan dan strategi untuk mencapai tujuan Data Mining berdasarkan informasi serta insight yang telah diperoleh [19].

b. Data Understanding

Tahap *Data Understanding* merupakan upaya dapat memahami menyeluruh mengenai data yang tersedia dalam proyek *Data Mining*, yang mencakup pengumpulan, deskripsi, eksplorasi pola, serta penilaian kualitas data[19].

c. Data Preparation

Pada fase Data *Preparation* bertujuan guna menyusun dataset akhir dari data mentah. Proses ini mencakup sejumlah tahapan, seperti pembersihan data guna mengatasi nilai yang hilang atau tidak valid, seleksi data dengan memilih atribut dan *record* yang dibutuhkan, serta mengubah data agar cocok untuk analisis. Output-nya menjadi input tahap pemodelan lebih lanjut [19].

d. Modelling

Tahap *modeling* diterapkan untuk mengekstraksi informasi berharga dari data yang telah dipersiapkan. Dalam tahap ini, metode *Data Mining* seperti pemodelan statistik dan *machine learning* digunakan sesuai dengan tujuan proyek serta karakteristik data. Model yang dikembangkan bertujuan guna menemukan pola dan tren, serta memprediksi atau mengklasifikasi data untuk menunjang keputusan Biasanya, proses ini melibatkan pengujian beberapa model untuk menentukan pendekatan terbaik yang paling relevan dengan data serta tujuan analisis [19].

e. Evaluation

Fase Evaluasi berfokus pada penilaian terhadap model ataupun teknik analisis yang diterapkan, untuk memastikan kualitas serta efektivitasnya. Proses evaluasi ini melibatkan penggunaan metrik kinerja guna mengidentifikasi kekuatan atau kelemahan model, sehingga perbaikan bisa dilakukan apabila diperlukan, agar solusi yang diterapkan berdampak maksimal bagi organisasi. [19].

f. Deployment

Setelah semua model selesai dibuat, diuji serta dievaluasi menerapkan datavalidasi, langkah berikutnya adalah tahap *deployment*, yang meliputi penyusunan laporan guna untuk memperlihat semua hasil yang didapatkan[19].

2.3.2 *Naïve Bayes Classifer (NBC)*

NBC ialah suatu algoritma yang diterapkan guna menentukan nilai probabilitas tertinggi yang kemudian digunakan untuk pengklasifikasian dan kemudian diuji dalam suatu kategori yang sesuai. Klasifikasi pada Naïve Bayes seringkali menjadi suatu pilihan dalam proses Data Mining dikarenakan penggunaannya yang relatif mudah dan pengimplementasiannya yang sederhana serta tingkat efektivitas yang cukup tinggi[20].

NBC ialah metode klasifikasi yang berbasiskan pada probabilitas serta Teorema Bayesian yang berasumsi bahwa setiap variabel *X* merupakan variabel bebas (*independence*) yang berartikan kebedaradaan suatu atribut (Variabel) tidak mempunyai keterkaitan dengan keberadaan atribut lainnya. Rumus perhitungan *Naïve Bayes Classifer* ditunjukkan pada rumus 2.1.

$$P(A\,|B) = \frac{P(B\,|A) * P(A)}{P(B)}$$

Rumus 2. 1Perhitungan Naïve Bayes Classifer (NBC)

Keterangan :

$P(A \mid B)$: Probabilitas bersyarat A yang diberikan oleh B

$P(B \mid A)$: Probabilitas bersyarat B yang diberikan oleh A

$P(A)$: Probabilitas kejadian A

$P(B)$: Probabilitas kejadian B

2.3.3 SVM (*Support Vector Machine*)

SVM atau biasa disebut juga sebagai *Support Vector Machine* ialah sebuah algoritma *machine learning* yang diterapkan guna mengklasifikasi dan regresi. Algoritma SVM termasuk ke dalam kategori *supervised learning*, yang berartikan bahwa algoritma SVM memerlukan data berlabel untuk melatih modelnya dan melakukan

prediksi pada sebuah data baru.

SVM sendiri merupakan sebuah algoritma yang sangat efektif dalam melakukan sebuah klasifikasi dan dikenal karena kemampuannya menangani kumpulan data yang kompleks[21].

2.3.4 *Random Forest*

Random Forest ialah algoritma pembelajaran *ensemble* yang mencakup sejumlah pohon keputusan yang dibangun menggunakan teknik *bagging* dan pemilihan fitur yang acak. Model ini mempunyai kelebihan yaitu dalam mengurangi *overfitting* yang sering terjadi pada pohon keputusan tunggal dan dapat meningkatkan akurasi prediksi[22]. Algoritma ini bekerja dengan menggabungkan beberapa hasil dari sejumlah pohon keputusan untuk menentukan keputusan akhir melalui voting(klasifikasi) atau *averaging*(regresi)[23].

2.3.5 *Confusion Matrix*

Confusion Matrix ialah sebuah tabel yang diterapkan guna mengevaluasi hasil dari kinerja klasifikasi, terutama dalam metode machine learning. Tabel ini akan memperlihatkan perbandingan dari hasil prediksi dengan label sebenarnya dari data, sehingga dapat diperlihatkan jumlah prediksi yang benar serta salah. *Confusion Matrix* mempunyai 4 istilah dalam representasi hasil klasifikasinya yaitu TP (*True positive*), TN *(True Negative*), FP (*False Positive)*, FN(*False Negative*). [24]

TP(*True Positive)* : terjadi saat data yang bernilai positif telah berhasil diklasifikasikan dengan tepat sebagai data positif

FP*(False Positive)* : muncul ketika data yang sebenarnya bernilai negatif, tetapi model salah dalam mengklasifikasikannya sebagai data positif.

FN*(False Negative)* : kondisi dimana ketika data seharusnya positif tetapi diprediksi sebagai data negatif.

TN*(True Negative)* : data yang bernilai negatif berhasil dikenali dengan benar sebagai data negatif.

Hasil dari *Confusion Matrix* diterapkan guna menghitung berbagai metrik evaluasi seperti *accuracy, precision, recall,* serta *F1-Score* melalui persamaan yang

telah ditentukan dengan tabel 2.2 yakni:

Tabel 2. 2Confusion Matrix

	True	False
True *(Positive)*	TP *(True Positive)*	FP *(False Positive)*
False *(Negative)*	FN *(False Negative)*	TN *(True Negative)*

2.3.6 *Accuracy*

Accuracy dapat menilai apakah pengklasifikasian model terolah dengan benar dan akurat. *Accuracy* ialah acuan guna memprediksi hasil positif, negatif, serta netral dari semua isi data. Rumus perhitungan *accuracy* ditunjukkan pada rumus 2.2. [24]

$$Accuracy = \frac{TN + TP}{TN + FP + FN + TP}$$

Rumus 2. 2Perhitungan Accuracy

2.3.7 *Recall*

Recall menunjukkan kemampuan model dalam mengenali informasi yang benar. Metrik ini dapat menggambarkan bagaimana presentase data positif yang berhasil di identifikasi dengan benar dan baik oleh model dibandingkan dengan total data positif yang ada. Rumus perhitungan *recall* ditunjukkan pada rumus 2.3. [24] :

$$Recall = \frac{TP}{TP + FN}$$

Rumus 2. 3Perhitungan Recall

2.3.8 *Precision*

Precision mengukur tingkat ketepatan antara data masukan dan *output* prediksi model. Metrik ini juga dapat dikenal sebagai rasio dari prediksi yang diklasifikasikan sebagai positif. Rumus perhitungan *precision* ditunjukkan pada rumus 2.4.[24] :

$$Precision : \frac{TP}{TP + FP}$$

Rumus 2. 4Perhitungan Precision

2.3.9 *F1-Score*

F1-Score memperlihatkan bagaimana nilai tengah dari presisi dan *recall score* dapat memprediksi hasil positif palsu serta negatif palsu. Rumus perhitungan *F1-Score* ditunjukkan pada rumus 2.5.[24]:

$$F1 - score = F1 = score = 2 \times \frac{precision \times recall}{precision + recall}$$

Rumus 2. 5Perhitungan F1-Score

2.3.10 *Text Mining*

Pada tahap ini, dijalankan proses ekstraksi teks dari berbagai dokumen sebagai bagian dari teknik *Text Mining*, dengan tujuan memperoleh informasi yang relevan sesuai kebutuhan analisis. Data yang dianalisis dalam *Data Mining* bisa berasal dari berbagai sumber, termasuk dari *web*. *Text Mining* sendiri menjadi salah satu cabang dari *data mining* yang berfokus pada pengolahan dan analisis dokumen teks menggunakan berbagai metode seperti klasifikasi atau kategorisasi[25]. Dalam konteks analisis sentimen, *Text Mining* digunakan untuk mengenali dan mengevaluasi sentimen yang terdapat dalam teks, yang positif, negatif, maupun netral-seperti yang terdapat dalam opini, kritik, ulasan, dan komentar. Bila diterapkan dengan baik, proses *Text Mining* dapat menghasilkan informasi yang bernilai dan mendukung individu atau organisasi dalam mengambil keputusan strategis bagi kepentingan bisnis.

2.3.10.1 *Data Preprocessing*

Data Preprocessing merupakan proses yang sangat penting dalam penerapan *Text Mining*. Proses ini bertujuan untuk mengubah teks mentah yang tidak terstruktur menjadi bentuk yang lebih rapih dan siap dianalisis. Tahapan dalam *text preprocessing* mencakup berbagai proses seperti *case folding*, tokenasi, *stemming*, serta penghapusan kata-kata umum(*stopword removal*)[25].

2.3.10.2 *Case Folding*

Case Folding merupakan salah satu tahap dalam pemrosesan teks yang bertujuan untuk menyamakan format penulisan dengan mengonversi seluruh huruf kapital menjadi huruf kecil. Selain itu, proses ini juga menghapus tanda baca dan karakter yang bukan merupakan huruf atau angka. Dengan menerapkan *case folding*, teks akan menjadi lebih konsisten sehingga memudahkan proses pencarian, analisis, maupun perbandingan. Prosedur ini menghilangkan elemen-elemen seperti huruf kapital, simbol, dan karakter khusus lainnya yang tidak berkaitan dengan isi utama teks. Sebagai contoh, kalimat "Bosen BA Artis Korea Terus" setelah melalui proses *case folding* akan berubah menjadi "bosen ba artis korea terus".

2.3.10.3 *Tokenizing*

Dalam tahap *tokenizing*, teks dibagi menjadi bagian-bagian kecil yang disebut sebagai token. Proses ini juga mencakup penghilangan simbol, angka, serta karakter lain yang tidak berkontribusi pada analisis kata. Tujuannya adalah untuk menyiapkan data teks agar dapat lebih mudah diolah oleh model. Setelah teks dipisahkan ke dalam bentuk token, proses lanjutan seperti pengolahan kata atau pengkodean dapat dilakukan dengan lebih efisien.

2.3.10.4 *Stopword Removal*

Pada tahap ini, hanya kata-kata yang dianggap penting yang akan dipertahankan dari hasil *tokenizing*, sementara kata-kata yang kurang relevan akan dihapus. Proses *stopword removal* ini umumnya dilakukan dengan dua pendekatan, yaitu :

1. *Stoplist*

   Stoplist merupakan kumpulan kata yang dinilai tidak memiliki pengaruh besar terhadap konteks analisis, seperti pemrosesan bahasa alami atau analisis teks. Kata-kata dalam daftar ini akan dikenali dan dihilangka dari dataset agar tidak mengganggu proses analisis selanjutnya. Menghapus kata-kata yang tidak esensial dapat membantu meningkatkan efisiensi serta ketepatan dalam tahap pengolahan data berikutnya.

2. *Wordlist*

Kumpulan kata signifikan dalam konteks tertentu dikenal sebagai *wordlist.* Penyusunannya dilakukan dengan memilih kata-kata kunci yang dipertahankan selama proses analisis ataupun pengolahan data. Pemilihan kata pada wordlist disesuaikan dengan tujuan penelitian, sementara kata yang tidak relevan dihapus. Penggunaan *wordlist* hanya memastikan bahwa hanya informasi penting yang diproses lebih lanjut dan hasil analisis menjadi lebih fokus dan efisien[25].

2.3.10.5 *Stemming*

Pada tahap ini, kata-kata akan diubah ke bentuk dasarnya selaras pada kaidah dalam Bahasa Indonesia. Proses ini dijalankan dengan menggunakan *library* yang terdapat dalam python yakni Sastrawi. Teknik ini berfungsi untuk menghilangkan afiks-afiks seperti awalan dan akhiran sehingga setiap kata dapat direduksi ke bentuk dasarnya.

2.3.11 SMOTE

SMOTE (*Synthetic Minority Over-Sampling Technique*) merupakan salah satu metode penyeimbangan data (*data balancing*) yang digunakan untuk menangani masalah klasifikasi dengan distribusi data yang tidak seimbang (*imbalanced classification*), yaitu ketika salah satu kelas memiliki jumlah data yang jauh lebih sedikit dibandingkan kelas lainnya. Ketidakseimbangan data ini seringkali menyebabkan model *machine learning* menjadi bias terdahap kelas mayoritas[26].

SMOTE bekerja dengan membuat sampel baru secara sintetis untuk kelas minoritas, bukan dengan melakukan duplikasi data secara langsung. Teknik ini dilakukan interpolasi antara satu titik data minoritas dengan beberapa tetangga terdekatnya. Sampel sintetis dibuat pada segmen garis antara titik tersebut dan tetangganya di ruang fitur[26].

2.4 Tools dan Software Penelitian

2.4.1 *Python*

Python merupakan sebuah bahasa pemrogramman yang dirancang dengan penekanan pada pembacaan kode. Bahasa ini memadukan sintaks yang jelas dengan

kemampuan yang kuat, sehingga sering digunakan sebagai bahasa scripting. Penggunaannya yang sangat luas dan mendukung pengembangan berbagai jenis aplikasi, serta dapat dijalankan di berbagai sistem operasi[27].

Dalam melakukan analisis data, python banyak dipilih karena sifatnya yang fleksibel dan didukung oleh banyak *library* populer seperti *Pandas, NumPy*, dan *Matplotlib*. Bahasa ini memudahkan analisis untuk menjalankan berbagai aktivitas, mulai dari pemrosesan data hingga pembuatan visualisasi. Sintaks python yang cenderung sederhana dapat membuatnya lebih mudah dipahami dan digunakan oleh banyak praktisi dari berbagai bidang. Selain itu, python juga unggul dalam menangani berbagai macam tipe data, baik yang terstruktur maupun tidak, sehingga sangat ideal untuk keperluan eksplorasi dan pemrosesan data secara efektif.

2.4.2 Google Colab

Google Colab adalah layanan komputasi berbasis *cloud* yang dikembangkan oleh Google, yang mendukung berbagai pustaka *machine learning* seperti *TensorFlow, NumPy, Pandas*, dan *Matplotlib.* Salah satu kelebihan utamanya adalah kemudahan dalam menampilkan visualisasi data tanpa harus menginstal perangkat lunak tambahan di komputer. Tak hanya untuk visualisasi, Google Colab juga memungkinkan pengguna melakukan *scraping data*, pemrosesan, hingga analisis data secara langsung di platform tersebut. Hal ini menjadikannya sangat ideal untuk digunakan dalam kegiatan penelitian maupun pengembangan di bidang data *science* dan kecerdasan buatan (AI). Selain itu, *platform* ini juga mendukung berbagai format file, termasuk *Jupyter Notebook*, sehingga memudahkan dalam membuat, menjalankan, dan membagikan *notebook* kepada pengguna lain [27].