BAB II
TINJAUAN PUSTAKA DAN LANDASAN TEORI

2.1 Tinjauan Pustaka

Penelitian terdahulu menjadi pedoman bagi peneliti untuk melakukan dan menyusun penelitian. Penelitian terdahulu dapat menjadi acuan bagi peneliti untuk memahami keterkaitan antara penelitian tersebut dengan penelitian yang akan dilakukan. Selain itu, penelitian terdahulu juga dapat menjadi dasar untuk menghindari duplikasi penelitian.

Pada penelitian[12] yang menerapkan metode *Naïve Bayes* untuk melakukan analisis sentimen terhadap pengambilalihan Taman Mini Indonesia Indah oleh pemerintah. Permasalahan yang diangkat adalah untuk melihat bagaimana perubahan pengelolaan yang dilakukan pemerintah terhadap TMII sebagai aset negara yang memiliki nilai keekonomian, sosial budaya, dan beragam nilai di dalamnya. Hasil penelitiannya menunjukkan bahwa algoritma *Naïve Bayes* memiliki akurasi sebesar 82.74% dan nilai AUC sebesar 0.500. Selain itu, penelitian ini juga menerapkan algoritma *Support Vector Machine* (SVM).

Penelitian sebelumnya[14] yang bertujuan untuk menganalisis sentimen terhadap aplikasi Ruang Guru dengan menggunakan komentar pengguna X. Metode analisis sentimen yang digunakan melibatkan algoritma *Naïve Bayes* (NB), *Support Vector Machine* (SVM), dan *K-Nearest Neighbour* (K-NN). Selain itu, dilibatkan juga feature selection menggunakan algoritma *Particle Swarm Optimization* (PSO). Hasil pengujian menunjukkan bahwa algoritma PSO berbasis SVM memberikan hasil optimal dengan akurasi 78,55% dan nilai *Area Under Curve* (AUC) 0,853. .

Pada penelitian[15] yang memiliki tujuan mengevaluasi sentimen terhadap perokok dan membedakan antara emosi positif dan negatif dalam konteks pembahasan merokok di Indonesia. Tiga metode klasifikasi, yaitu *Naïve Bayes* (NB*), Support Vector Machine* (SVM), dan *Logistic Regression*,

digunakan untuk menganalisis sentimen dari data yang dikumpulkan. Hasil penelitian ini menunjukkan bahwa sebanyak 40,25% pengguna X setuju dengan keberadaan perokok di Indonesia, sedangkan 59,74% tidak setuju. Metode *Naïve Bayes* memberikan hasil terbaik dengan tingkat akurasi sebesar 62,1%, menggunakan pembagian data latih sebesar 60% dan data uji sebesar 40%.

Pada penelitian[16] yang bertujuan untuk menganalisis sentimen masyarakat terkait proyek pengembangan di Pulau Rinca dan Pulau Komodo. Penelitian ini menggunakan metode analisis sentimen berdasarkan pada teknik *machine learning* dan pemrosesan bahasa alami. Metode yang digunakan dalam penelitian ini melibatkan penggunaan model Doc2Vec yang terdiri dari *distributed model* dan *distributed bag of words*, serta penggunaan *support vector machines* dan *logistic regression* sebagai *classifier*. Hasil dari penelitian ini menunjukkan bahwa sebagian besar sentimen masyarakat cenderung menentang pengembangan di Pulau Rinca. Dalam penelitian ini, setiap kombinasi model dan *classifie*r memiliki tingkat akurasi di atas 75%, yang menunjukkan bahwa hampir semua sentimen masyarakat menentang pengembangan di Pulau Rinca. PV-DBOW dengan Regresi Logistik memiliki akurasi 0.86858974359, sedangkan PV-DM dengan SVM memiliki akurasi 0.80750538769.

Penelitian sebelumnya[17] yang bertujuan untuk melakukan analisis perbandingan berbagai model *machine learning* dan *deep learning* dalam klasifikasi teks bahasa Inggris dan Bangla. Studi ini berfokus pada analisis sentimen komentar dari situs e-commerce Bengali populer, "DARAZ," yang terdiri dari ulasan Bangla dan terjemahan Inggris. Penelitian ini mengimplementasikan tujuh model *machine learning* dan *deep learning*, seperti *Long Short-Term Memory* (LSTM), *Bidirectional LSTM* (Bi-LSTM), *Convolutional 1D* (Conv1D), dan gabungan Conv1D-LSTM. Teknik praproses diterapkan pada kumpulan teks yang dimodifikasi untuk meningkatkan akurasi model. Hasil penelitian menjelaskan bahwa model *Support Vector Machine* (SVM) menunjukkan kinerja yang lebih unggul dibandingkan model lain,

dengan akurasi 82,56% untuk analisis sentimen teks bahasa Inggris dan 86,43% untuk analisis sentimen teks Bangla menggunakan algoritma porter *stemming*. Selain itu, Model Berbasis Bi-LSTM menunjukkan kinerja terbaik di antara model *deep learning*, dengan akurasi 78,10% untuk teks bahasa Inggris dan 83,72% untuk teks Bangla menggunakan porter *stemming*.

Pada penelitian[18] yang bertujuan membandingkan akurasi model analisis sentimen menggunakan Word2Vec dan *FastText* pada ulasan hotel berbahasa Indonesia. Kedua model *Word embedding* ini berdasarkan *Skip-gram* diujicobakan pada dataset TripAdvisor. Parameter-parameter seperti jumlah fitur, minimum kata, *thread paralel*, dan ukuran jendela konteks diatur sama untuk kedua model. Hasil penelitian menunjukkan bahwa baik *FastText* maupun Word2Vec berhasil meningkatkan akurasi pada model *Random Forest* dan *Extra Tree*, dengan *FastText* mencapai kinerja lebih baik, meningkatkan akurasi sebesar 8% dari baseline (*Decision Tree* 85%) menjadi 93% dengan 100 estimator. Penelitian ini menyoroti keunggulan *FastText* dalam konteks analisis sentimen pada ulasan hotel bahasa Indonesia.

Penelitian sebelumnya[19] yang bertujuan untuk melakukan analisis sentimen opini masyarakat terhadap program Kampus Merdeka di X sebagai respons terhadap kebijakan pendidikan tinggi baru yang diperkenalkan oleh Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi Republik Indonesia sejak Januari 2019. Pengumpulan data dilakukan melalui situs web vicinitas.io dengan memantau hashtag #kampusmerdeka dan #mbkm pada *tweet* dan *retweet* pengguna X selama November 2021 hingga Maret 2022. Analisis dilakukan terhadap 501 *tweet* menggunakan algoritma klasifikasi *Naïve Bayes*, dengan hasil menunjukkan 272 sentimen positif dan 229 sentimen negatif. Metode ini dievaluasi dengan akurasi rata-rata 60%, presisi 64%, *recall* 58%, dan *f1-score* 58%. Hasil penelitian ini memberikan gambaran sentimen dominan seperti "kampus," "merdeka," dan "mbkm" pada sentimen positif, serta kata-kata seperti "kampus," "uang," dan "saku" pada sentimen negatif, memberikan wawasan yang berharga untuk evaluasi algoritma, kinerja program, dan tanggapan masyarakat terhadap inisiatif pendidikan tersebut.

Pada penelitian[20] yang memiliki tujuan menganalisis sentimen tanggapan netizen terhadap berita resesi 2023 di X dan menentukan apakah tanggapan tersebut bersifat positif atau negatif. Metode yang digunakan dalam penelitian ini menggunakan *Support Vector Machine*(SVM). Hasil penelitian menunjukkan akurasi sebesar 98,67%, dengan recall sebesar 99,33% untuk sentimen positif dan 98,00% untuk sentimen negatif. Presisi sebesar 98,03% untuk sentimen positif dan 99,32% untuk sentimen negatif . Selain itu, penelitian ini bertujuan untuk memahami dan menentukan persentase akurasi dari setiap kelas sentimen.

Penelitian[21] yang bertujuan untuk membandingkan metode *Long Short Term Memory* (LSTM) dengan metode Naïve Bayes dalam analisis sentimen terhadap kebijakan new normal. Metode LSTM dipilih karena memiliki kemampuan untuk menyimpan informasi jangka panjang, membaca, dan memperbarui informasi sebelumnya. Sedangkan metode *Naïve Bayes* dipilih karena telah terbukti memiliki kinerja yang baik dalam beberapa penelitian sebelumnya. Hasil penelitian menunjukkan bahwa metode LSTM dan *Naïve Bayes* memiliki kinerja yang hampir sama dalam analisis sentimen terhadap kebijakan new normal. Kedua metode tersebut memiliki akurasi, presisi, *recall*, dan *f-measure* yang tinggi. Namun, metode LSTM memiliki waktu eksekusi yang lebih lama dibandingkan dengan metode Naïve Bayes.

Dalam penelitian[22] yang bertujuan untuk menganalisis dampak dari teknik ekstraksi fitur pada analisis sentimen terhadap dataset SS-Tweet. Masalah yang ingin dipecahkan adalah menentukan teknik ekstraksi fitur mana yang memberikan kinerja terbaik dalam analisis sentimen terhadap dataset tersebut. Metode penelitian melibatkan penggunaan enam teknik pra-pemrosesan dan ekstraksi fitur, diikuti dengan penerapan enam algoritma klasifikasi dan evaluasi terhadap empat parameter kinerja. Hasil penelitian menunjukkan bahwa penggunaan teknik ekstraksi fitur TF-IDF pada tingkat kata (*Term Frequency-Inverse Document Frequency*) memberikan kinerja 3-4% lebih tinggi daripada menggunakan fitur N-Gram dalam analisis sentimen terhadap dataset SS-Tweet. Metode klasifikasi yang memiliki akurasi tinggi dalam analisis sentimen data X adalah *Logistic Regression*.

Metode ini memberikan hasil yang lebih baik dibandingkan dengan metode lain seperti *Random Forest*, *Decision Tre*e, *Naïve Bayes*, SVM, dan KNN.

Penelitian ini memiliki beberapa perbedaan dengan penelitian sebelumnya. Perbedaan pertama terletak pada dataset yang digunakan, dimana penelitian ini memanfaatkan komentar dari media sosial X. Pada penelitian-penelitian sebelumnya menggunakan sumber data yang berbeda, dan ada juga yang menggunakan sumber yang sama namun dengan fokus objek yang berbeda. Perbedaan kedua terletak pada pilihan algoritma, penelitian ini menggunakan algoritma *Support Vector Machine* (SVM).

Tabel 2.1 Penelitian Sebelumnya

No	Judul	Penulis	Masalah	Metode	Hasil Penelitian
1	Analisis Sentimen Opini Publik Terhadap Pengambil Alihan TMII Oleh Pemerintah Dengan Algoritma *Naïve Bayes*(2023).	Ika Amelia, Adinda Mutiara, Imam Santoso.	Sentimen masyarakat terhadap perubahan pengelolaan yang dilakukan pemerintah terhadap TMII.	Algoritma *Naïve Bayes*.	Algoritma *Naïve Bayes* memiliki akurasi sebesar 82.74% dan nilai AUC sebesar 0.500 . Selain itu, penelitian ini juga menggunakan algoritma *Support Vector Machine* (SVM).
2	Analisis Sentimen Aplikasi Ruang Guru Di Twitter Menggunakan Algoritma Klasifikasi(2020)	Angelina Puput Giovani, Ardiansyah, Tuti Haryanti, Laela Kurniawati, Windu Gata	Analisis sentimen terhadap aplikasi Ruang Guru berdasarkan data X tersebut.	*Naïve Bayes* (NB), *Support Vector Machine* (SVM), dan *K-Nearest Neighbour* (K-NN)	Hasil penelitian ini menunjukkan bahwa algoritma SVM dengan optimasi PSO memberikan hasil optimal dalam klasifikasi sentimen, dengan akurasi sebesar 78,55% dan nilai *Area Under Curve* (AUC) sebesar 0,853. Dengan demikian, penelitian ini berhasil menemukan algoritma yang efektif dan terbaik dalam mengklasifikasikan komentar positif dan negatif terkait dengan aplikasi Ruang Guru
3	Analisa Sentimen Pengguna Sosial Media Twitter Terhadap Perokok di Indonesia(2023)	Dewi Setiyawati dan Nuri Cahyono	Mengevaluasi sentimen terhadap perokok dan membedakan antara emosi positif dan negatif dalam konteks pembahasan merokok di Indonesia.	*Naïve Bayes* (NB), *Support Vector Machine* (SVM), dan *Logistic Regression*	Hasil penelitian menunjukkan bahwa sebanyak 40,25% pengguna X setuju dengan keberadaan perokok di Indonesia, sedangkan 59,74% tidak setuju. Metode *Naïve Bayes* memberikan hasil terbaik dengan tingkat akurasi sebesar

No	Judul	Penulis	Masalah	Metode	Hasil Penelitian
					62,1%, menggunakan pembagian data latih sebesar 60% dan data uji sebesar 40%.
4	*Sentiment analysis of Twitter data related to Rinca Island development using Doc2Vec and SVM and logistic regression as classifier(2022)*	Tirta Hema Jaya Hidayat, Yova Ruldeviyani, Achmad Rizki Aditama, Gusti Raditia Madya, Ade Wija Nugraha, Muhammad Wijaya Adisaputra	Analisis sentimen masyarakat terkait proyek pengembangan di Pulau Rinca dan Pulau Komodo.	*Doc2Vec, Support Vector Machine, Logistic Regression.*	Hasil dari penelitian ini menunjukkan bahwa sebagian besar sentimen masyarakat cenderung menentang pengembangan di Pulau Rinca. Dalam penelitian ini, setiap kombinasi model dan classifier memiliki tingkat akurasi di atas 75%, yang menunjukkan bahwa hampir semua sentimen masyarakat menentang pengembangan di Pulau Rinca. PV-DBOW dengan Regresi Logistik memiliki akurasi 0.86858974359, sedangkan PV-DM dengan SVM memiliki akurasi 0.80750538769.
5	*Sentiment analysis in multilingual context: Comparative analysis of machine learning and hybrid deep learning models(2023)*	Rajesh Kumar Das, Mirajul Islam, Md Mahmudul Hasan, Sultana Razia, Mocksidul Hassan, Sharun Akter Khushbu	Analisis perbandingan berbagai model *machine learning* dan deep learning dalam klasifikasi teks bahasa Inggris dan Bangla	*Logistic Regression, Decision Tree, Random Forest, Multi Naïve Bayes*, KNN, SVM, SGD, LSTM *Based Model*, Bi-LSTM *Based Model*, Conv1D	Hasil penelitian menunjukkan bahwa model Support Vector Machine (SVM) menunjukkan kinerja yang lebih unggul dibandingkan model lain, dengan akurasi 82,56% untuk analisis sentimen teks bahasa Inggris dan 86,43% untuk analisis sentimen teks Bangla menggunakan algoritma porter

No	Judul	Penulis	Masalah	Metode	Hasil Penelitian
				Based Model, Conv1D–LSTM *Based Model*	stemming. Selain itu, Model Berbasis Bi-LSTM menunjukkan kinerja terbaik di antara model *deep learning*, dengan akurasi 78,10% untuk teks bahasa Inggris dan 83,72% untuk teks Bangla menggunakan porter stemming.
6	*The Accuracy Comparison Between Word2Vec and FastText On Sentiment Analysis of Hotel Reviews(2022)*	Siti Khomsah, Rima Dias Ramadhani dan Sena Wijayanto	Membandingkan akurasi model analisis sentimen menggunakan Word2Vec dan FastText pada ulasan hotel berbahasa Indonesia	*Random Forest*, *Decision Tree*	Hasil penelitian menunjukkan bahwa baik FastText maupun Word2Vec berhasil meningkatkan akurasi pada model Random Forest dan Extra Tree, dengan FastText mencapai kinerja lebih baik, meningkatkan akurasi sebesar 8% dari baseline (Decision Tree 85%) menjadi 93% dengan 100 estimator. Penelitian ini menyoroti keunggulan FastText dalam konteks analisis sentimen pada ulasan hotel bahasa Indonesia
7	Analisis Sentimen Terhadap Program Kampus Merdeka Menggunakan Algoritma *Naïve Bayes* Classifier Di Twitter(2023)	Elisa Febriyani dan Herny Februariyanti	Menganalisis sentimen opini publik terhadap program Kampus Merdeka di X sebagai respons terhadap kebijakan pendidikan tinggi baru yang diperkenalkan oleh	*Naïve Bayes Classifier*	Berdasarkan hasil penelitian, sistem berhasil mengklasifikasikan 272 hasil sentimen positif dan 229 pendapat sentimen negatif dengan akurasi rata-rata sebesar 60%, presisi 64%, *recall* 58%, dan f1-score 58%. Visualisasi

No	Judul	Penulis	Masalah	Metode	Hasil Penelitian
			Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi Republik Indonesia sejak Januari 2019		hasil ditampilkan dalam bentuk word cloud dengan kata-kata dominan pada sentimen positif antara lain "kampus," "merdeka," "mbkm," dan "program," sedangkan pada sentimen negatif antara lain "kampus," "uang," "saku," dan "konversi."
8	Analisis Sentimen Dengan Algoritma SVM Dalam Tanggapan Netizen Terhadap Berita Resesi 2023(2023)	Dadang Iskandar Mulyana dan Nesti Lutfianti	Sentimen analisis terhadap isu potensialnya resesi pada tahun 2023 di Indonesia melalui platform media sosial X	*Support Vector Machine* (SVM)	Hasil pengujian model menunjukkan tingkat akurasi sebesar 98,67%, dengan nilai *Recall* Positif 99,33%, *Recall* Negatif 98,00%, *Precision* Positif 98,03%, dan *Precision* Negatif 99,32%. Proses analisis sentimen berhasil memprediksi dengan akurat opini masyarakat terhadap isu resesi 2023 di X, memberikan gambaran yang kuat terkait sentimen positif dan negatif yang muncul di kalangan netizen
9	Analisis Perbandingan Algoritma LSTM dan *Naïve Bayes* untuk Analisis Sentimen(2022)	Auliya Rahman Isnain, Heni Sulistiani, Bagus Miftaq Hurohman, Andi Nurkholis, Styawati Styawati	membandingkan kinerja metode Long Short Therm Memory (LSTM) dengan *Naïve Bayes* terhadap analisis sentimen Kebijakan New Normal	*Long Short Therm Memory* (LSTM) dan *Naïve Bayes*	Hasil penelitian menunjukkan bahwa metode LSTM memiliki kinerja yang lebih baik bila dibandingkan dengan *Naïve Bayes*. Metode LSTM menghasilkan nilai akurasi, presisi dan *recall* sebesar 83.33%. Sedangkan metode

No	Judul	Penulis	Masalah	Metode	Hasil Penelitian
					Naïve Bayes memiliki nilai akurasi, presisi dan *recall* sebesar 82%.
10	*The Impact of Features Extraction on the Sentiment Analysis* (2019)	Ravinder Ahuja, Aarkhasa Chug, Shruti Kohli, Shaurya Gupta, dan Pratyush Ahuja	Menentukan teknik ekstraksi fitur mana yang memberikan kinerja terbaik dalam analisis sentimen terhadap dataset SS-Tweet.	*Decision Tree, Support Vector Machine, K-Nearest Neighbor, Random Forest, Logistic Regression, dan Naïve Bayes*	Hasil penelitian menunjukkan bahwa penggunaan teknik ekstraksi fitur TF-IDF pada tingkat kata (Term Frequency-Inverse Document Frequency) memberikan kinerja 3-4% lebih tinggi daripada menggunakan fitur N-Gram dalam analisis sentimen terhadap dataset SS-Tweet. Metode klasifikasi yang memiliki akurasi tinggi dalam analisis sentimen data X adalah *Logistic Regression*. Metode ini memberikan hasil yang lebih baik dibandingkan dengan metode lain seperti *Random Forest*, *Decision Tree*, *Naïve Bayes*, SVM, dan KNN

2.2 Landasan Teori

2.2.1 Text Mining

Text mining juga dikenal sebagai penambangan data teks atau analisis teks, adalah proses untuk mendapatkan informasi berkualitas tinggi dari teks. Ini melibatkan penemuan informasi baru yang sebelumnya tidak diketahui dengan mengekstraksi informasi secara otomatis dari berbagai sumber teks. Proses ini biasanya mencakup interpretasi informasi yang terkumpul dan dapat digunakan untuk klasifikasi prediktif, pengisian basis data, atau pembuatan indeks pencarian berdasarkan informasi yang diekstraksi[23].

Text mining melibatkan analisis data teks yang tidak terstruktur untuk mengekstraksi pola dan pengetahuan yang signifikan guna mendukung pengambilan keputusan. Proses ini mencakup sejumlah teknik seperti ekstraksi informasi, pengambilan informasi, pemrosesan bahasa alami, pengelompokan, kategorisasi, visualisasi, dan rangkuman teks. Tujuan utama dari *text mining* adalah mengungkap informasi berharga dari kumpulan dokumen besar yang ditulis dalam bahasa alami. Hal ini sangat penting dalam konteks penerbitan dengan basis data informasi yang besar yang memerlukan indeksasi untuk pencarian, serta dalam disiplin ilmiah di mana informasi yang sangat spesifik sering ditemukan dalam teks tertulis[24].

2.2.2 Analisis Sentimen

Analisis sentimen atau juga dikenal sebagai penambangan opini, adalah teknik pemrosesan bahasa alami (NLP) yang digunakan untuk menentukan apakah data bersifat positif, negatif, atau netral[25]. Penggunaan analisis sentimen seringkali dilakukan untuk menilai pendapat atau perasaan yang terdapat dalam teks, seperti tinjauan produk, konten media sosial, atau hasil survei pelanggan. Teknik ini memungkinkan perusahaan untuk memahami bagaimana pelanggan merespons produk atau layanan mereka, serta untuk mengidentifikasi area yang perlu diperbaiki atau ditingkatkan. Analisis sentimen juga dapat digunakan untuk mengelompokkan teks berdasarkan sentimen yang terkandung di dalamnya, serta untuk mengaitkan sentimen

dengan fitur spesifik dari produk atau layanan, dalam apa yang dikenal sebagai analisis sentimen berbasis aspek[26].

Analisis sentimen menggunakan teknologi *Natural Language Processing*(NLP) dan *machine learning*(ML) untuk melatih perangkat lunak dalam menganalisis dan menginterpretasikan teks. Perangkat lunak ini dapat menggunakan pendekatan *aspect-based*, *machine learning*, atau kombinasi keduanya yang disebut hibrida. Pendekatan *aspect-based* mengklasifikasikan kata kunci dalam teks menggunakan leksikon kata-kata positif dan negatif. Sementara itu, pendekatan *machine learning* menggunakan algoritma seperti *regresi linier, Naive Bayes, Support Vector Machine*, dan *deep learning* untuk mengidentifikasi sentimen dalam teks. Pendekatan hibrida menggabungkan kedua metode untuk meningkatkan akurasi dan kecepatan.

Dalam konteks *text mining* penerapan analisis sentimen menjadi aspek penting yang memungkinkan organisasi memahami serta menanggapi pandangan dan perasaan yang terdapat dalam teks. Hal ini berguna untuk meningkatkan proses pengambilan keputusan dan melakukan perbaikan pada produk atau layanan. Meskipun menghadapi kendala seperti bahasa informal dan perubahan dalam konteks, analisis sentimen tetap berperan krusial dalam pemahaman opini masyarakat dan menjaga kualitas hubungan antara organisasi dan konsumen. Evaluasi yang teliti terhadap hasil analisis sentimen menjadi kunci untuk memastikan akurasi interpretasi dan memberikan dasar yang solid bagi pengambilan keputusan yang efektif[27].

2.2.3 Klasifikasi Teks

Klasifikasi teks adalah suatu proses di dalam analisis data dan pemrosesan bahasa alami yang bertujuan untuk mengelompokkan atau mengategorikan dokumen atau potongan teks ke dalam kategori atau label tertentu berdasarkan isinya[28]. Tujuan utama dari klasifikasi teks adalah untuk mengotomatiskan dan menyederhanakan tugas-tugas seperti identifikasi topik, deteksi sentimen, atau pengelompokan berita. Klasifikasi teks memiliki berbagai manfaat, seperti memungkinkan analisis data yang cepat dan efektif, pengawasan keadaan darurat, deteksi respons negatif atau

darurat, dan kategorisasi data yang tidak terstruktur ke dalam kelompok. Teknik ini juga dapat digunakan untuk mengembangkan alat klasifikasi teks berbasis kecerdasan buatan yang efektif, dapat diskalakan, dan hemat biaya.

Proses klasifikasi teks melibatkan penggunaan algoritma pembelajaran mesin yang dilatih dengan menggunakan data pelatihan yang telah dikategorikan sebelumnya. Algoritma ini belajar dari pola-pola atau fitur-fitur yang muncul dalam teks untuk dapat mengklasifikasikan teks baru ke dalam kategori yang sesuai[29]. Beberapa langkah utama dalam klasifikasi teks melibatkan pemrosesan teks, ektraksi fitur, *machine learning* ataupun *deep learning*, evaluasi dan validasi. Keterampilan utama dalam klasifikasi teks melibatkan pemilihan fitur yang tepat dan pemilihan model *machine learning* atau *deep learning* yang sesuai dengan tugas klasifikasi yang spesifik[30].

2.2.4 Text Preprocessing

Preprocessing (pra-pemrosesan) adalah serangkaian langkah atau teknik yang diterapkan pada data sebelum data tersebut digunakan dalam suatu analisis atau model. Tujuan utama dari pra-pemrosesan adalah untuk membersihkan, mengorganisir, dan mempersiapkan data agar dapat diolah dengan lebih efektif oleh algoritma atau model yang akan digunakan[31]. Tahapan dalam *preprocessing* terdiri dari *case folding*, *cleansing*, *tokenizing*, *stopword*, dan *stemming*[20].

a. *Case Folding*

   Pada proses casefolding dilakukan untuk mengubah semua karakter huruf dalam sebuah teks menjadi huruf kecil, dengan tujuan untuk membuat perbandingan dan pencarian teks menjadi lebih konsisten dan tidak terpengaruh oleh perbedaan huruf besar dan huruf kecil.

b. *Cleansing*

   Cleansing atau pembersihan data adalah langkah untuk membersihkan data yang baru saja diambil dari X. Pada tahap ini, data yang diperoleh melalui *crawling* dapat mengandung beberapa teks yang tidak relevan, seperti hastag, URL, tanda @, *retweet*, dan elemen lainnya. Oleh karena

itu, diperlukan proses pembersihan untuk menghilangkan elemen-elemen tersebut.

c. *Tokenizing*
Proses *tokenizing* proses memecah suatu teks atau dokumen menjadi unit-unit yang lebih kecil, yang disebut sebagai token. Token dapat berupa kata, frasa, atau simbol, tergantung pada tingkat granularitas yang diinginkan.

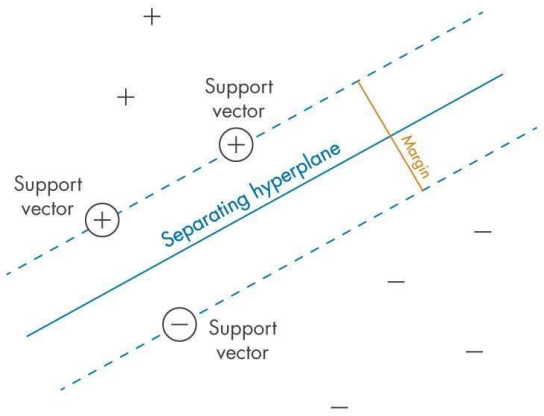
d. *Stopword*
Pada fase ini, dilakukan ekstraksi kata-kata kunci dari hasil tokenisasi. Proses ini melibatkan penghapusan kata-kata yang memiliki sedikit makna atau tidak relevan, seperti contohnya "yang", "di", "ke", “cukup”, “membuat” dan juga kata hubung atau konjungsi seperti dan", "atau", "tetapi", "sebab", "karena. Pada proses stopword menngunakan kamus stopword berbahasa Indonesia.

e. *Stemming*
Stemming adalah proses pembersihan suatu kata dengan menghapus imbuhan sehingga diperoleh bentuk dasar atau akar kata, *stemming* ini menggunakan *library* dari Sastrawi

2.2.5 Algoritma Support Vector Machine

Support Vector Machine (SVM) merupakan suatu algoritma dalam *machine learning* yang digunakan untuk tugas klasifikasi maupun regresi. SVM melakukan analisis data dengan tujuan menemukan *hyperplane* yang dapat memisahkan antara kelas data yang berbeda[32]. SVM berguna untuk menganalisis data yang kompleks yang tidak dapat dipisahkan oleh garis lurus sederhana, dan mengubah data masukan menjadi ruang fitur dimensi yang lebih tinggi untuk menemukan pemisahan linear atau untuk mengklasifikasikan set data secara lebih efektif. Tujuan dari SVM adalah menemukan *hyperplane* yang memaksimalkan margin, yaitu jarak antara *hyperplane* dan titik data terdekat dari setiap kelas[33]. Titik-titik data ini disebut vektor pendukung. SVM menggunakan teknik yang disebut fungsi kernel untuk memetakan data ke dalam ruang dimensi yang lebih tinggi, di

mana data menjadi linear dapat dipisahkan. Hal ini memungkinkan SVM untuk menangani hubungan non-linear antara titik-titik data.



Gambar 2.1 Support Vector Machine

Pada gambar 2.1 menunjukkan sebuah diagram *hyperplane* pemisah dan vektor pendukung. *Hyperplane* pemisah adalah sebuah garis atau bidang yang memisahkan dua kelas data. Vektor pendukung adalah data poin yang terletak paling dekat dengan *hyperplane* pemisah. Gambar tersebut menunjukkan tiga vektor pendukung, dua dari kelas positif dan satu dari kelas negatif. Margin adalah jarak antara *hyperplane* pemisah dan vektor pendukung terdekat dari setiap kelas. SVM bertujuan untuk menemukan *hyperplane* pemisah yang memaksimalkan margin. Margin yang lebih besar berarti bahwa *hyperplane* pemisah lebih meyakinkan dalam memisahkan dua kelas data.

Rumus untuk SVM menggunakan *Support Vector Classifier* (SVC) dengan kernel linear dapat dibagi menjadi dua bagian, dapat dilihat pada persamaan (2.1) dan (2.2)

1. Rumus untuk *Hyperplane*

$$w \cdot x + b = 0 \qquad (2.1)[34]$$

Keterangan :

w: Vektor bobot yang menentukan arah *hyperplane*

x: Vektor data

b: Bias term, yang menentukan posisi *hyperplane* pada sumbu y

2. Rumus untuk fungsi objektif

$$\min \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{n} \xi_i \quad (2.2)[34]$$

Keterangan:

w: vektor bobot

ξ_i: variabel slack yang mengukur seberapa jauh sampel ke-i berada dari margin yang diharapkan

C: parameter regularisasi yang mengontrol optimasi antara margin dan kesalahan klasifikasi

Proses klasifikasi data SVM dengan sebuah objek data x dapat diformulasikan pada persamaan 2.3

$$f\big(\Phi(x)\big) = w \cdot \Phi(x) + b \quad (2.3)\ [34]$$

$$= \sum_{i=1, xi \in SV}^{n} a_i \; y_i \, \Phi(x) \cdot \Phi(x_i) + b$$

$$= \sum_{i=1, xi \in SV}^{n} a_i \; y_i \, K(x, x_i) + b$$

dimana SV adalah objek objek data pada himpunan data latih yang terpilih sebagai *support vector*

Berikut langkah-langkah untuk algoritma pembelajaran SVM untuk menemukan *hyperplane* optimum berbasis metode sekuensial dapa dilihat pada persamaan 2.4

1. Initizalization, $a_i = 0$ (2.4)[34]

   Hitung matiks $D_{ij} = y_i \; y_j \, (K\big(x_i, x_j\big) + \lambda^2$

2. Lakukan tiga langkah di bawah ini untuk i = 1,2,...,l

   a. $E_i = \sum_{j=1}^{l} a_j \, D_{Ij}$

   b. $\delta a_i = \min \{\max[\gamma(1 - E_1), -a_i], C - a_i\}$

   c. $a_i = a_i + \delta a_i$

3. Kembali ke langkah 2 sampai α konvergen

Pada algoritma persamaan 2.4 , x adalah vektor masukan yang diperluas dengan skalar λ dan D adalah matriks yang dimodifikasi, γ berfungsi untuk mengontrol kecepatan belajar. Suatu kondisi konvergen didefinisikan jika perubahan α relatif kecil.

2.2.6 *Crawling*

Crawling adalah proses penjelajahan internet secara otomatis untuk mengumpulkan data dari berbagai halaman web. Proses ini melibatkan mengunjungi halaman web, mengikuti tautan, dan mengindeks konten yang ditemukan[35]. Fungsi dari crawling adalah untuk mengumpulkan dan mengelola informasi yang tersedia di internet secara otomatis dan efisien. Data hasil crawling tersebut kemudian bisa dimanfaatkan untuk berbagai keperluan.

Proses web crawling dimulai dengan menyediakan URL awal atau daftar URL yang ingin diambil, dan bot mengakses halaman web pertama dari daftar tersebut menggunakan protokol HTTP atau HTTPS. Setelah mengambil halaman web, bot menganalisis HTML untuk mengekstrak informasi seperti teks, gambar, dan tautan. Selanjutnya, spider mencari tautan dalam halaman yang baru saja diambil, membangun suatu jaringan tautan yang memungkinkannya menjelajahi situs web secara menyeluruh. Informasi yang ditemukan disimpan atau diindeks untuk penggunaan di masa mendatang, seperti membangun indeks mesin pencari atau mengumpulkan data statistik. Proses ini berulang dengan mengambil halaman-halaman terhubung dari tautan-tautan yang ditemukan. Web crawling memiliki berbagai aplikasi, termasuk pengindeksan mesin pencari, pengumpulan data analisis, dan pemantauan perubahan pada situs web.

2.2.7 *Tweet Harvest*

Tweet Harvest adalah sebuah *command line tool* yang menggunakan *Playwrigh*t untuk mengambil *tweet* dari hasil pencarian X berdasarkan kata kunci dan rentang tanggal tertentu. *Tweet* yang diambil disimpan dalam file CSV. Alat ini memerlukan akun X yang valid dan *Access Token* yang diperoleh dengan login ke X di browser dan mengekstrak *authorization token*.

Proses penggunaan *Tweet* Harvest cukup sederhana. Pengguna hanya perlu menjalankan perintah "*npx tweet-harvest@latest"* dan menekan enter. Alat ini akan membuka browser Chromium, menavigasi ke halaman pencarian X, memasukkan parameter pencarian yang telah ditentukan, dan mengambil *tweet* yang dihasilkan. *Tweet* tersebut akan disimpan dalam file CSV di direktori bernama "*tweet*s-data" di direktori kerja saat itu[36].

2.2.8 Term Frequency - Inverse Document Frequency(TF-IDF)

TF-IDF adalah metode yang digunakan untuk mengukur seberapa penting suatu kata dalam suatu dokumen. Metode ini didasarkan pada dua konsep, yaitu *term frequency* (TF) dan *inverse document frequency* (IDF)[20]. *Term frequency* (TF) adalah jumlah kemunculan suatu kata dalam sebuah dokumen. Semakin sering suatu kata muncul, semakin tinggi nilai TF. *Inverse document frequency* (IDF) adalah ukuran seberapa umum suatu kata digunakan dalam seluruh dokumen. Semakin umum suatu kata digunakan, semakin rendah nilai IDF. Nilai TF-IDF adalah hasil perkalian antara TF dan IDF. Nilai TF-IDF yang tinggi menunjukkan bahwa suatu kata penting dalam suatu dokumen, karena kata tersebut muncul sering dalam dokumen tersebut dan tidak umum digunakan dalam dokumen lain[32].

Perhitungan bobot pada setiap dokumen dapat dilakukan dengan menggunakan persamaan (2.5).

$$W_{i,j} = tf_{i,j} \times \log(\frac{n}{df_i}) \quad \textbf{(2.5)}[19]$$

Keterangan :

$W_{i,j}$: Bobot kata i pada dokumen j

$tf_{i,j}$: Frekuensi kata i pada dokumen j

n : Jumlah dokumen

df_i : Jumlah dokumen yang terdapat kata i

2.2.9 Confusion Matrix

Confusion Matrix adalah suatu tabel yang digunakan pada evaluasi klasifikasi model dalam *machine learning*. *Confusion Matrix* merupakan tabel khusus yang memungkinkan visualisasi kinerja suatu algoritma,

biasanya algoritma *supervised learning*. Ini digunakan untuk mengukur kinerja model klasifikasi, yang bertujuan untuk memprediksi label kategoris untuk setiap contoh input[37]. *Confusion Matrix* menampilkan jumlah *true positives* (TP), *true negatives* (TN), *false positives* (FP), dan *false negatives* (FN) yang dihasilkan oleh model pada data uji. Untuk klasifikasi biner, matriks akan berbentuk tabel 2x2, sedangkan untuk klasifikasi multi-kelas, bentuk matriks akan sama dengan jumlah kelas, yaitu n x n. Pada Tabel 2.2, menunjukkan kinerja model klasifikasi pada sekumpulan data uji dengan nilai sebenarnya yang diketahui.

Tabel 2.2 Confusion Matrix

Confusion matrix		***Actual***	
		Positive	***negative***
Predicted	***Positive***	***True Positive***	***False Negative***
	Negative	***False Negatif***	***True Positive***

Terdapat beberapa rumus untuk menghitung performa klasifikasi dari confusion matrix yaitu *accuracy, precision, recall* dan *f1 score*[38]. *Accuracy* adalah ukuran seberapa baik model secara keseluruhan dalam memprediksi kelas yang benar. Rumus untuk menghitung *accuracy* terdapat pada persamaan(2.6).

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \qquad \textbf{(2.6)}$$

Precision adalah ukuran seberapa baik model memprediksi kelas positif yang sebenarnya adalah positif. Semakin tinggi *precision*, semakin baik model dalam memprediksi kelas positif yang benar. Rumus menghitung *precision* terdapat pada persamaan(2.7).

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \qquad \textbf{(2.7)}$$

Recall adalah ukuran seberapa baik model mendeteksi semua kelas positif yang sebenarnya adalah positif. Semakin tinggi *recall*, semakin baik model dalam mendeteksi kelas positif yang benar. Rumus menghitung *recall* terdapat pada persamaan(2.8).

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \qquad \textbf{(2.8)}$$

F1-score adalah kombinasi dari presisi dan recall. Semakin tinggi F1-score, semakin baik model dalam memprediksi kelas yang benar dan mendeteksi kelas positif yang benar. Rumus menghitung F1-score terdapat pada persamaan(2.9).

$$\text{F1} - \text{score} = 2 \text{ x} \frac{\text{Precision x Recall}}{\text{Precision+ Recall}} \qquad \textbf{(2.9)}$$