BAB II
TINJAUAN PUSTAKA

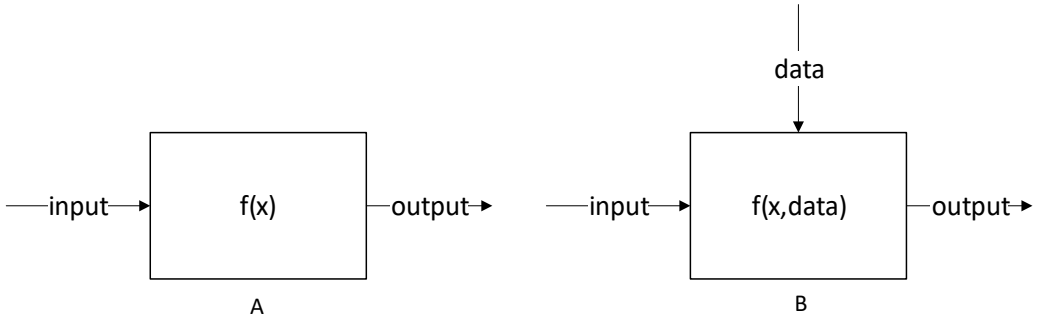
II.1 *Machine Learning*

Machine learning ialah salah satu bagian dari kecerdasan kecerdasan buatan. Pada *machine learning* ini dibuat agar sistem atau mesin dibuat menjadi cerdas seperti manusia pada umumnya. Namun pada hal ini, kecerdasan dibuat dengan cara proses pembelajaran dan pelatihan terlebih dahulu, sebelum sistem tersebut melakukan pada dunia nyata. Dengan demikian, segala aktifitasnya akan mudah dikenali, dipahami, dan dikerjakan dengan baik, efektifdan efisien Kelebihan dan kekurangan (Rachmad, 2020).

Terdapat beberapa macam definisi mengenai *machine learning*. Menurut *Expert System* pada tahun 2017, *machine learning* adalah aplikasi atau bagian dari kecerdasan buatan yang membuat sistem memiliki kemampuan belajar secara otomatis dan meningkatkan kemampuannya berdasarkan pengalaman tanpa diprogram secara eksplisit. Dalam hal ini, program komputer tidak ditulis secara statis. Fokus *machine learning* terdapat pada pengembangan program komputer yang dapat mengakses data dan belajar dari data tersebut. *Machine learning* juga dapat didefinisikan sebagai algoritma yang bertujuan untuk menemukan dan mengaplikasikan pola-pola di dalam data. Algoritma pada *machine learning* menggunakan teknik-teknik statistik untuk menemukan pola-pola tersebut. Seringkali data yang dicari polanya berukuran besar. Data yang diolah juga tidak hanya teks, tetapi dapat juga berupa gambar, audio, video, atau aktivitas-aktivitas

pengguna selama berselancar atau mengakses internet. Menurut definisi yang Iain, *machine learning* adalah ilmu yang memungkinkan komputer berperilaku seperti manusia, di mana komputer dapat meningkatkan pemahamannya melalui pengalaman atau dengan berjalannya waktu secara otomatis. Kemampuan belajar tersebut diperoleh dengan mengakses data dan informasi secara terus-menerus (Kusuma, 2020).

Berdasarkan paparan di atas, terdapat beberapa kata kunci, yaitu: belajar, kecerdasan buatan, kode program tidak eksplisit, data masukan. Berdasarkan kata-kata kunci tersebut, *machine learning* dapat didefinisikan sebagai suatu algoritma atau program komputer yang dapat membuat sistem menjadi cerdas dengan mempelajari data- data yang tersedia di mana algoritma atau program tersebut tidak didefinisikan secara eksplisit. Dengan demikian, terdapat perbedaan mendasar antara algoritma *machine learning* dengan algoritma pemrograman umum. Pada algoritma pemrograman umum, algoritma ditulis secara eksplisit. Ilustrasi dapat dilihat pada Gambar 2.1 di mana Gambar 2.1A adalah diagram algoritma umum dan Gambar 2.1B adalah diagram algoritma *machine learning* (Kusuma, 2020).



Gambar II.1 Diagram Algoritma Umum dan Algoritma *Machine Learning*

Penjelasan Gambar 2.1 adalah sebagai berikut. Pada Gambar 2.1A, terdapat suatu proses di mana *output* merupakan luaran fungsi f(x) di mana x adalah masukan fungsi. Adapun pada Gambar 2.1B, *output* merupakan luaran fungsi f di mana yang menjadi masukan adalah *input* dan data (Kusuma, 2020).

II.2 Feature Selection

Pengertian *feature selection* adalah teknik menyeleksi atüibut pada sebuah dataset, hingga terpilih attribut-atribut terpenting sesuai karakteristik objek tertentu (Mesran dkk, 2023). *Feature selection* adalah identifikasi dan pemilihan subset fitur yang relevan dari dataset asli dengan berfokus pada fitur yang paling informatif (R. F. Putra dkk, 2023). *Feature selection* adalah salah satu cara untuk mereduksi dimensi dataset yang besar dan meningkatkan keakuratan. Menurut *Marcin Blachnik* tentang *Comparison of Various Feature Selection Methods in Application to Prototype Best Rules*. *Marcin Blachnik* menjelaskan bahwa *feature selection* berperan memilih subset yang tepat dari set fitur asli, karena tidak semua fitur atau atribut relevan dengan masalah. Bahkan beberapa dari fitur atau atribut tersebut mengganggu dan dapat mengurangi akurasi. *Noisy features* atau fitur yang tidak terpakai tersebut harus dihapus untuk meningkatkan akurasi (Indriyanto, 2021).

Tujuan utama dari *feature selection* adalah untuk menghapus atribut yang tidak terpakai atau yang tidak terkumpul dengan baik. *Feature selection* umumnya lebih sulit untuk masalah yang tidak diawasi, seperti pengelompokan, di mana kriteria validasi eksternal, seperti label, tidak tersedia untuk *feature selection*. Secara intuitif, masalah *feature selection* terkait erat dengan masalah penentuan kecenderungan *clustering inheren* dari satu set fitur. Metode *feature selection*

menentukan subset fitur yang memaksimalkan kecenderungan pengelompokan yang mendasarinya (Purwati & Kurniawan, 2021). Ada dua *class* model utama untuk melakukan *feature selection* diantaranya (Purwati & Kurniawan, 2021):

- Model *Filter* : Dalam hal ini, skor dikaitkan dengan setiap fitur dengan menggunakan kriteria berbasis kesamaan. Kriteria ini pada dasarnya adalah filter yang memberikan kondisi yang tepat untuk penghapusan fitur. Poin data yang tidak memenuhi Skor yang disyaratkan akan dihapus dari pertimbangan. Dalam beberapa kasus, model ini dapat mengukur kualitas subset fitur sebagai kombinasi, bukan fitur tunggal. Model semacam itu lebih kuat karena secara implisit memperhitungkan dampak tambahan dari penambahan fitur ke yang lain.
- Model *wrapper* : algoritma pengelompokan digunakan untuk mengevaluasi kualitas subset fitur. Ini kemudian digunakan untuk memperbaiki subset fitur di mana pengelompokan dilakukan. Ini adalah pendekatan yang berulang secara alami di mana pilihan fitur yang baik bergantung pada *cluster* dan sebaliknya. Fitur yang dipilih biasanya akan tergantung pada metodologi tertentu yang digunakan untuk pengelompokan. Meskipun ini mungkin tampak seperti kerugian, kenyataannya adalah bahwa metode pengelompokan yang berbeda dapat bekerja lebih baik dengan set fitur yang berbeda. Oleh karena itu, metodologi ini juga dapat mengoptimalkan pemilihan fitur ke teknik *clustering* tertentu. Di sisi lain, sifat keinformatifan yang melekat dari fitur-fitur tertentu terkadang tidak dapat direfleksikan oleh pendekatan ini karena dampak dari metodologi *clustering* tertentu.

Perbedaan utama antara model *filter* dan model *wrapper* adalah model yang sebelumnya dapat dilakukan murni sebagai fase praproses, sedangkan model terakhir diintegrasikan langsung ke dalam proses pengelompokan (Purwati & Kurniawan, 2021).

II.3 Algoritma *Support Vector Machine*

Support Vector Machine (SVM) merupakan algoritma *machine learning* dengan prinsip pengklasifikasian data dengan menggunakan garis pemisah terbaik (*hyperplan*) yang mampu mengoptimalkan margin (jarak tegak lurus amatan positif dan negatif terdekat terhadap garis *hyperplan*) melalui proses minimalisasi menggunakan teknik *Lagrange*. Amatan positif dalam pengertian ini merupakan amatan yang memiliki selisih positif terhadap garis *hyperplan*, sedangkan amatan negatif merupakan amatan yang memiliki selisih negatif terhadap garis *hyperplane* (Nursiyono, 2023).

Salah satu algoritma pembelajaran mesin yang digunakan dalam proses klasifikasi dan regresi dikenal sebagai *Support Vector Machine* (SVM). Metode SVM adalah salah satu cara untuk mengklasifikasikan opini menjadi beberapa bagian, untuk membangun model menggunakan data yang telah disediakan dengan mencari *hyperplane* yang memberikan kemungkinan pemisahan terbaik antara dua kelas data. *Hyperplane* yang dimaksud dapat dianggap sebagai garis, bidang, atau hyperplane karena membagi data menjadi dua kategori berbeda. Untuk mengurangi kemungkinan klasifikasi yang salah saat diterapkan pada data yang belum pernah dilihat sebelumnya, SVM mengoptimalkan margin dengan meningkatkan jarak

antara *hyperplane* dan titik data yang mewakili setiap kelas yang paling dekat dengannya (Amrullah dkk, 2023).

Mesin vektor dukungan (SVM) dapat diterapkan pada data *linier* dan *non-linier*, dan menggunakan teknik kernel untuk mengubah data menjadi dimensi yang lebih tinggi sehingga dapat dibagikan Oleh *hyperplanes*. SVM dapat diterapkan untuk berbagai macam masalah, termasuk kategorisasi teks, identifikasi gambar, dan prediksi harga pasar. Algoritma SVM sering digunakan dalam analisis sentimen sebagai alat untuk mengkategorikan teks ke dalam kategori positif, negatif, atau netral. Ini dimungkinkan Oleh kapasitas algoritma yang kuat untuk menangani masalah klasifikasi yang rumit(Amrullah dkk, 2023).

Cara terbaik untuk memperkenalkan *Support Vector Machines* (SVMs) adalah dengan mempertimbangkan tugas sederhana klasifikasi biner. Banyak masalah dunia nyata melibatkan prediksi atas dua kelas. Dengan demikian kami dapat ingin memprediksi apakah nilai tukar mata uang akan bergerak naik atau turun, tergantung pada data ekonomi, atau apakah klien harus diberi pinjaman berdasarkan informasi keuangan pribadi atau tidak. Sebuah SVM adalah mesin pembelajaran abstrak yang akan belajar dari kumpulan data pelatihan dan mencoba untuk menggeneralisasi dan buat prediksi yang benar pada data baru. Untuk data pelatihan memiliki seperangkat vektor input, dilambangkan Xi, dengan setiap vektor input memiliki sejumlah fitur komponen. Vektor input ini dipasangkan dengan (i =1,.....,m) (Campbell & Ying, 2022).

II.3.1 Konsep dasar *support vector machine*

Setelah memahami definisi dan kelebihan serta kelemahan dari SVM, selanjutnya kita akan membahas mengenai fungsi kernel dan *margin hyperplane* dalam SVM (Amrullah dkk, 2023).

- Fungsi Kernel

  Fungsi kernel dalam SVM merupakan sebuah teknik untuk mentransformasi data input ke dalam bentuk yang lebih mudah diproses. Fungsi kernel ini memungkinkan SVM untuk melakukan klasifikasi pada data yang tidak dapat dipisahkan secara linear di dalam ruang input. Dengan menggunakan fungsi kernel, data input dapat diubah ke dalam bentuk yang memungkinkan SVM untuk mencari *hyperplane* yang dapat memisahkan data dengan sebaik-baiknya. Fungsi kernel ini dapat berbentuk *linear*, *polynomial*, *sigmoid*, atau *radial basis function* (RBF). Masing-masing jenis kernel memiliki karakteristik dan kegunaannya tersendiri dalam menangani berbagai jenis masalah klasifikasi. Dalam SVM, pemilihan jenis kernel yang tepat sangat penting untuk memaksimalkan performa model. Oleh karena itu, pemilihan jenis kernel harus dilakukan dengan cermat berdasarkan karakteristik dari data yang akan diklasifikasikan.

- Margin dan Hyperplane

  Margin dan hyperplane merupakan dua konsep penting dalam SVM yang sangat berhubungan dengan pemisahan antara dua kelas data yang akan diklasifikasikan. Margin adalah jarak antara hyperplane dengan titik terdekat dari kelas data. Hyperplane sendiri adalah sebuah garis atau bidang

yang berfungsi sebagai batas pemisah antara dua kelas data. Pada kasus SVM dengan dua dimensi, hyperplane berupa garis yang membagi bidang menjadi dua bagian yang masing-masing berisi data dari kelas yang berbeda. Sedangkan pada kasus SVM dengan tiga dimensi, hyperplane berupa bidang yang membagi ruang menjadi dua bagian. Dalam SVM, tujuan utama adalah menemukan *hyperplane* yang dapat memisahkan dua kelas data secara optimal dengan maksimum margin, yaitu jarak antara hyperplane dengan titik terdekat dari kelas data. Margin yang maksimum dianggap sebagai solusi terbaik karena hyperplane yang dihasilkan akan memiliki kesalahan prediksi yang lebih rendah. Pencarian hyperplane optimal dan maksimum margin dilakukan melalui proses pelatihan SVM dengan menggunakan data latih yang telah diberi label kelas. SVM akan mencari hyperplane yang memaksimalkan margin dengan memperhitungkan sejumlah parameter yang disebut dengan parameter C dan parameter gamma pada fungsi kernel. Parameter-parameter tersebut dapat diatur dan dioptimalkan untuk memaksimalkan performa SVM dalam melakukan klasifikasi pada data baru. pada SVM *Hyperplane* diposisikan sejauh mungkin dari *support vectors* yang paling dekat dengan setiap kelas. Dalam *hyperplane* terdapat posisi optimal yang didefinisikan sebagai berikut (Amrullah dkk, 2023) :

$$h.d + C = 0 \qquad (1)$$

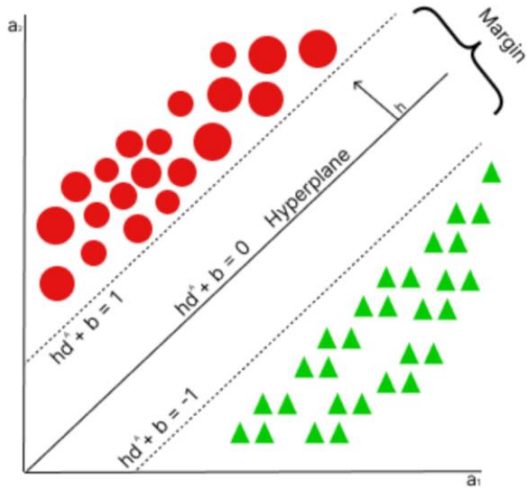
Penjelasan rumus (1), *h* diartikan sebagai *weight vector*, d diartikan sebagai *input feature vector*, dan *C* sebagai bias. h dan *C* akan dapat memenuhi ketidaksetaraan berikut untuk semua komponen set pelatihan. Setelah

hyperplane optimal didapat, hyperplane dapat digunakan untuk melakukan klasifikasi. Adapun rumus *hypotesis function* sebagai berikut (Amrullah dkk, 2023) :

$$h.d + C \geq +1 \; if \; y_k = 1 \qquad (2)$$

$$h.d + C \leq -1 \; if \; y_k = -1 \qquad (3)$$

Saat melatih model SVM, tujuannya adalah untuk menemukan *hyperplane* yang dapat memisahkan data secara akurat. Adapun ilustrasi *hyperplane* optimal sebagai berikut :



Gambar II.2 *Hyperlane Optimal*

Untuk mendapatkan hyperplane yang akurat maka perlu dilakukan perbandingan *hyperplane*, adapun rumus *compare hyperplanes* sebagai berikut (Amrullah dkk, 2023):

$$F = \frac{\min}{i=1.....A} fi(hd + C) \qquad (4)$$

Penjelasan pada rumus (4), F diartikan sebagai *functional margin of the dataset*, h diartikan sebagai *weight vector*, d diartikan sebagai *input feature*

vector, dan C sebagai bias. Setelah mendapatkan *metrics compare hyperlanes,* maka tahapan selanjutnya mencari nilai *weight vector* dan bias. Adapun rumus mencari nilai *weight vector* dan biasnya sebagai berikut (Amrullah dkk, 2023):

$$h = \sum_{i=1}^{A} ai \quad KiJi \tag{5}$$

Penjelasan pada rumus *weight vector,* h diartikan sebagai *weight vector*, *a* diartikan sebagai *Lagrangian Multiplier Value*, K dan J merupakan Vektor.

$$C = \frac{1}{V}\sum_{i=1}^{V}(Ki - h.d) \tag{6}$$

Penjelasan pada rumus (6), C diartikan sebagai bias, V diartikan sebagai *support vectors*, h diartikan sebagai *weight vector* dan K merupakan vektor(Amrullah dkk, 2023).

II.3.2 *Support vector machine* dalam analisis sentimen

Implementasi SVM dalam analisis sentimen bertujuan untuk mengklasifikasikan teks atau kalimat menjadi positif atau negatif. Langkah-langkah implementasi SVM dalam analisis sentimen adalah sebagai berikut (Amrullah dkk, 2023):

1. Pengumpulan dan *Preprocessing* Data: Data teks atau kalimat dikumpulkan dan diproses terlebih dahulu. Proses preprocessing meliputi penghapusan karakter khusus, tokenisasi, dan *stemming*. Tujuan dari preprocessing adalah untuk menghasilkan data yang lebih mudah diproses dan akurat.

2. Pembagian Data: Setelah data diproses, data dibagi menjadi dua bagian yaitu data *training* dan data *testing*. Data *training* digunakan untuk melatih model SVM, sedangkan data *testing* digunakan untuk menguji model.
3. Pemilihan Fungsi *Kernel*: Fungsi *kernel* digunakan untuk mengubah data *input* ke dalam bentuk yang dapat diproses oleh SVM. Beberapa jenis fungsi *kernel* yang biasa digunakan adalah *linear*, *polynomial*, dan *radial basis function* (RBF). Pemilihan fungsi kernel yang tepat akan mempengaruhi akurasi dari model SVM.
4. Pelatihan Model SVM: Setelah fungsi *kernel* dipilih, model SVM dilatih menggunakan data *training*. Model akan belajar untuk membedakan kalimat positif dan negatif berdasarkan fitur yang diberikan.

II.4 Algoritma *Chi-Square*

Algoritma *Chi-square* merupakan salah satu metode seleksi fitur *supervised* yang mampu menghilangkan fitur tanpa mengurangi tingkat akurasi yang dihasilkan. Algoritma *Chi-square* bertujuan untuk melihat ketergantungan suatu fitur terhadap kategori atau label kelasnya. Algoritma dari seleksi fitur algoritma *Chi-square* adalah mengurangi data yang besar menjadi lebih kecil sehingga proses yang dilakukan menjadi lebih cepat (Chairunnisa dkk, 2022).

Pengertian *Chi-square* adalah uji non parametik yang biasa digunakan yang sedang melakukan penelitian. Bukan hanya untuk penelitian dalam skala besar, para mahasiswa yang sedang mengerjakan skripsi pun pasti akan berhubungan dengan uji *Chi-square* tersebut karena umumnya terdapat pada pengolahan data yang menggunakan software statistik seperti SPSS atau SPLS. Algoritma *Chi-square*

disebut juga dengan Kai Kuadrat. Uji ini tidak bisa dilakukan begitu saja mengingat metode dan rumus yang cukup kompleks, sehingga untuk menerapkannya diperlukan bimbingan khusus dari yang sudah berpengalaman. Uji *Chi-square* biasanya digunakan untuk mengetahui hubungan dua variabel nominal kemudian mengukur kekuatan hubungan antara dua variabel yang dimaksud. Skala data kedua variabel adalah nominal. Apabila dari 2 variabel, ada 1 variabel dengan skala nominal maka dilakukan uji *Chi-square* dengan merujuk bahwa harus digunakan uji pada derajat yang terendah. Uji *Chi-square* bisa dilakukan hanya pada sampel berukuran besar. Uji ini dilakukan dengan mentabulasikan variabel ke dalam kategori-kategori lalu dihitung statistik *chi-square* nya (Usman & Akbar, 2020).

Menurut (Ari Wibowo, 2015) uji *chi-square* memeriksa perbedaan antara nilai yang diamati dan nilai yang diharapkan. *Chi-Square* menunjukkan atau dengan cara memeriksa hubungan antara dua variabel kategoris yang dapat dihitung dengan menggunakan frekuensi yang diamati yang diberikan dan frekuensi yang diharapkan. *Chi-Square* memiliki rumus :

$$X^2 = \sum \frac{(O_i - E_i)^2}{E_i} \qquad (7)$$

Dimana :

O_i = nilai yang diamati (nilai sebenarnya)

E_i = nilai yang diharapkan

Karakteristik *chi-square* memiliki nilai *chi-square* selalu positif, terdapat beberapa keluarga distribusi *chi-square*, distribusi *chi-square* dengan DK=1, 2, 3, dan seterusnya dan bentuk distribusi *chi-square* adalah menjulur positif. Fungsi *chi-square* untuk menguji hubungan atau pengaruh dua buah variabel nominal dan

mengukur kuatnya hubungan antara variabel yang satu dengan variabel nominal lainnya (C = *coefisien of contingency*). Pada dasarnya, jika Anda adalah seorang peneliti maka perlu rasanya untuk mempelajari rumus Chi Square harus anda pelajari agar dapat mengerti makna sesungguhnya dari uji chi square. Sehingga dapat memanfaatkan rumus tersebut dalam kegiatan analisis, terutama saat melakukan uji kumparatif pada data dengan skala data nominal (Usman & Akbar, 2020).

Dalam pengklasifikasian, *chi-square* adalah salah satu kategori supervised seleksi fitur yang mampu menghilangkan banyak fitur tanpa mengurangi tingkat akurasi dan *chi-square* juga banyak digunakan untuk menghasilkan kinerja yang bagus dalam keakuratan. Algoritma seleksi fitur *chi-square* dapat mereduksi data yang besar sehingga proses yang dibutuhkan menjadi cepat. Terdapat beberapa algoritma seleksi fitur seperti *document frequency*, *information gain*, *chi-square*, *term strength* dan *mutual information*. Pada penelitian yang dilakukan menunjukkan bahwa chi-square termasuk algoritma seleksi fitur efisien dan lebih baik (Rifqi Tsani dkk, 2020). Untuk itu *chi-square* tepat digunakan sebagai algoritma seleksi fitur pada analisis sentimen *twitter* dengan *support vector machine*.

II.4.1 Syarat uji *Chi-square*

Syarat umum uji *Chi Square* adalah frekuensi atau sampel yang digunakan besar, sebab ada beberapa syarat *Chi Square* dapat digunakan yaitu :

- Tidak ada cell dengan nilai frekuensi kenyataan atau disebut juga Actual Count (F0) sebesar 0 (Nol).

- Apabila bentuk tabel kontingensi 2 X 2, maka tidak boleh ada 1 cell saja yang memiliki frekuensi harapan atau disebut juga expected count (“Fh”) kurang dari 5.
- Apabila bentuk tabel lebih dari 2 x 2, misak 2 x 3, maka jumlah cell dengan frekuensi harapan yang kurang dari 5 tidak boleh lebih dari 20%.

II.5 *Twitter*

Twitter merupakan sebuah media sosial yang saat ini banyak digunakan oleh masyarakat Indonesia sebagai ajang eksistensi diri, khususnya pada kalangan mahasiswa. Banyak cara yang dapat dilakukan orang untuk dapat meningkatkan eksistensi dirinya di *Twitter*. Dengan menunjukkan penampilan profil serta gaya berbicaranya, dapat menunjukkan kepada pengikutnya mengenai kesan yang ingin ditampilkan di *Twitter* (Aulia Girnanfa & Susilo, 2022).

Twitter didirikan oleh 3 (tiga) orang, yaitu: *Jack Dorsey*, *Biz Stone*, dan *Evan Williams* pada bulan Maret tahun 2006, dan baru diluncurkan pada bulan Juli di tahun yang sama. *Twitter* adalah jejaring sosial dan *micro-blogging*, yang memfasilitasi, sebagai pengguna, dapat memberikan pembaharun informasi tentang diri sendiri, bisnis, dan lain sebagainya. Setiap menulis status pada Twitter, status tersebut disebut sebagai *Tweets*. Apabila jumlah *Tweets* Anda berjumlah 50 (lima puluh), maka sudah menulis status pada *Twitter* sebanyak 50 kali. *Tweets* merupakan penulisan teks berbasis 140 (seratus empat puluh) karakter. Jadi, jumlah maksimal karakter yang dituliskan sebagai status hanya terbatas pada jumlah maksimal 140 (seratus empat puluh) karakter. *Tweets* dapat ditampilkan pada profil akun atau dapat mengomentari status dari teman. Keistimewaan *Tweets* adalah

dapat mengirimkannya melalui Twitter via *internet, SMS* (*Short Message Service*), atau aplikasi-aplikasi yang lain. Tentunya, pengiriman *Tweets* melalui *SMS* akan berpengaruh pada penggunaan pulsa ponsel yang kita miliki (Jati Waloeyo, 2010).

Dampak positif dengan adanya *Twitter*, dapat menambah pertemanan, baik yang berasal dari dalam negeri maupun beberapa yang berasal dari luar negeri. *Twitter* juga dapat dijadikan sebagai sarana dalam memperluas pengetahuan bahasa. Selain itu *Twitter* membantu dalam memperluas wawasan seperti mendapatkan kosakata baru serta ilmu baru lainnya dan terdapat pengetahuan yang diperoleh dari *Twitter*. *Twitter* juga menjadi media yang digunakan sebagai sarana pembelajaran serta menambah pertemanan, dimana dalam menemuka informasi baru (Annisa Batu Bara dkk, 2022).

Dampak negatif dari *Twitter* media yang digunakan untuk menyebarkan informasi palsu (hoax) dan propaganda. Sejalan dengan berkembangnya teknologi di era saat ini membuat orang-orang menjadi mudah dalam memperoleh informasi yang mereka butuhkan. Terdapat beberapa pengguna *Twitter* sering menggunakan *Twitter* untuk berbagi informasi kepada pengguna lainnya serta ada juga yang memanfaatkannya untuk menyebarkan hoax. Hal ini dikarenakan masih ditemukannya akun *Twitter* yang belum terorganisir dengan baik yang membuat konten dewasa masih dapat diakses secara luas oleh siapapun. Dampak negatif lainnya yaitu adanya kebebasan dari *Twitter* digunakan untuk mengunggah konten dewasa (Annisa Batu Bara dkk, 2022).

II.6 Analisis Sentimen

Analisis sentimen merupakan ekstraksi dari polaritas dan subjektivitas dari orientasi semantik yang berhubungan dengan arti dari sebuah kata & teks atau polaritas frasa. Teknik pemrosesan bahasa alami yang disebut analisis sentimen digunakan untuk mengekstraksi dan menganalisis sentimen atau opini dari dokumen teks seperti berita, postingan media sosial, atau evaluasi produk. Analisis sentimen adalah contoh teknik pemrosesan bahasa alami. Sentimen analisis bertujuan untuk mendeteksi emosi yang tersirat dalam sebuah teks. Maka dari itu, Sentimen analisis dapat diterapkan ke pendapat dari suatu produk, layanan, organisasi & topik . Ide mendasar di balik analisis sentimen adalah menggunakan pemrosesan bahasa alami untuk menentukan perasaan yang disampaikan Oleh sepotong teks tertulis. Dalam analisis sentimen, ada tiga bentuk utama sentimen: positif, negatif, dan netral. Secara umum, sentimen positif dianggap paling umum. Sentimen positif dapat dipahami sebagai sikap yang menyenangkan atau kesukaan terhadap suatu subjek, sedangkan sentimen negatif mengacu pada pandangan yang tidak menyenangkan atau ketidaksukaan terhadap subjek (Amrullah dkk, 2023).

Di Sisi Iain, teks dengan sentimen netral adalah teks yang tidak secara jelas mengungkapkan sentimen positif atau negatif. Dokumen teks perlu diproses terlebih dahulu menggunakan teknik pemrosesan bahasa alami seperti *tokenization*, *stemming*, dan penghapusan *stop word* sebelum analisis sentimen dapat dilakukan pada dokumen tersebut. Setelah itu, algoritma klasifikasi, seperti *Naive Bayes* dan *Support Vector Machines* (SVM), atau algoritma *deep learning*, seperti *Recurrent Neural Networks* (RNN) atau *Convolutional Neural Networks*, digunakan untuk

mengekstrak sentimen atau opini dari dokumen teks. Beberapa paper telah dilakukan terkait dengan sentimen analisis. Mengumpulkan dan menganalisis data dari sumber publik seperti media sosial, forum diskusi, atau jajak pendapat *online* adalah jenis sumber publik yang dapat dimanfaatkan untuk analisis sentimen. Analisis sentimen dapat digunakan untuk mengukur sentimen atau opini publik terhadap topik tertentu. Hasil analisis sentimen dapat digunakan untuk membantu pemangku kepentingan dalam mengambil keputusan, seperti untuk meningkatkan layanan pelanggan, untuk mengetahui pendapat masyarakat tentang kebijakan pemerintah, atau untuk memantau reputasi merek atau produk (Amrullah dkk, 2023).

II.7 Penelitian Deskriptif

Jenis penelitian yang akan digunakan pada penelitian ini yaitu penelitian non-implementatif. Penelitian non-implementatif termasuk dalam kategori penelitian deskriptif. Menurut (Rusardi dkk, 2022) Penelitian non-implementatif merupakan sebuah tipe penelitian yang berfokus pada suatu investigasi fenomena atau situasi tertentu yang selanjutnya akan ditinjau secara ilmiah.

Penelitian deskriptif adalah metode penelitian yang bertujuan untuk menggambarkan atau mengidentifikasi fenomena atau karakteristik suatu populasi atau situasi. Penelitian deskriptif bertujuan untuk memberikan gambaran yang jelas dan sistematis mengenai fenomena yang sedang diteliti, tanpa melakukan manipulasi atau pengendalian variabel. Penelitian deskriptif biasanya dilakukan dengan mengumpulkan data dari subjek penelitian dalam kondisi alamiah atau yang sudah ada. Data yang dikumpulkan dapat berupa data kualitatif (misalnya

observasi, wawancara, atau studi kasus) atau data kuantitatif (misalnya survei, pengukuran, atau analisis statistik) (Syahza, 2021).

Tujuan penelitian deskriptif adalah untuk membuat penyanderaan atau gambaran secara sistematis, faktual dan akurat mengenai fakta-fakta dan sifat-sifat populasi atau daerah tertentu. Secara harfiah, ciri-ciri penelitian deskiriptif adalah penelitian yang bermaksud untuk membuat penyanderaan (deskripsi) mengenai situasi-situasi atau kejadian-kejadian. Dalam arti ini penelitian deskriptif adalah akumulasi data dasar dalam cara deskriptif semata-mata, tidak perlu mencari atau menerangkan saling hubungan, menguji hipotesis, membuat ramalan, atau mendapatkan makna dan implikasi, walaupun penelitian yang bertujuan untuk menemukan hal-hal tersebut dan mencakup juga metode-metode deskriptif. Tetapi para ahli dalam bidang penelitian tidak ada kesepakatan mengenai apa sebenarnya penelitian deskriptif itu. Sementara ahli memberikan arti penelitian deskriptif itu lebih luas, dan mencakup segala macam bentuk penelitian kecuali penelitian historis dan penelitian eksprimental (Syahza, 2021).

II.8 Penelitian Terdahulu

Adapun beberapa penelitian yang menjadi acuan dalam penelitian ini dapat dilihat pada tabel

Tabel II.1 Penelitian Terdahulu

No	Penelitian/ Tahun Terbit	Judul Penelitian	Persamaan	Perbedaan
1.	(Kencana dan Sibaroni 2019)	Klasifikasi *Sentiment Analysis* pada Review Buku Novel Berbahasa	Menggunakan *support*	Analisis sentimen

		Inggris dengan Menggunakan Metode *Support Vector Machine* (SVM)	*vector machine*	pada buku novel
2.	(Rifqi Tsani dkk 2020)	Analisis Sentimen Review Transportasi Menggunakan Algoritma *Support Vector Machine* Berbasis *Chi Square*	Menggunakan *support vector machine* berbasis *chi square*	Analisis sentimen *review* pada situs internet *Yelp*
3.	(Chairunnisa dkk 2022)	Klasifikasi Sentimen Ulasan Pengguna Aplikasi Peduli Lindungi di *Google Play* Menggunakan Algoritma *Support Vector Machine* dengan Seleksi Fitur *Chi-Square*	Menggunakan *support vector machine* dengan seleksi fitur *chi square*	Analisis sentimen ulasan pada aplikasi peduli lindungi
4.	(Agatha & Maria Polina, 2022)	Analisis Sentimen Terhadap Penggunaan *Marketplace* Di Indonesia Menggunakan Metode *Support Vector Machine*	Menggunakan *support vector machine*	Analisis sentimen *tweet* layanan masalah tokopedia dan shopee
5.	(Luthfiana dkk, 2020)	Implementasi Algoritma *Support Vector Machine* dan *Chi Square* untuk Analisis Sentimen *User Feedback* Aplikasi	Menggunakan *support vector machine* dan *Chi Square*	Analisis sentimen user *feedback* aplikasi

6.	(Yang dkk, 2018)	Enhanced Twitter Sentiment Analysis by Using Feature Selection and Combination	menggunakan *naive bayes multinominal* dengan seleksi *chi square*	Analisis sentimen *tweet* dengan kombinasi 6 algoritma

II.9 Kontibusi Penelitian

Penelitian sebelumnya oleh (Chairunnisa dkk, 2022) Klasifikasi Sentimen Ulasan Pengguna Aplikasi Peduli Lindungi di *Google Play* Menggunakan Algoritma *Support Vector Machine* dengan Seleksi Fitur *Chi-Square* berfokus pada analisis sentimen seluruh ulasan pengguna aplikasi peduli lindungi. Selain itu penelitian (Rifqi Tsani dkk, 2020) Analisis Sentimen Review Transportasi Menggunakan Algoritma *Support Vector Machine* Berbasis *Chi Square* berfokus pada menganalisa hasil sentimen ulasan pengguna pada situs *Yelp*. Dalam hal ini penulis melakukan penelitian ini berfokus pada analisis data *tweet* menggunakan *Support Vector Machine* (SVM) dan *Chi Square* terhadap seleksi fitur. Tahap awal pengumpulan data sentimen dari *Twitter* yang disebut dengan *dataset*. Dataset menjadi *input* utama pada proses *Chi Square*. *Chi Square* digunakan untuk menilai hubungan antara variabel. Variabel mengacu pada kata-kata atau fitur dalam *tweet*. Hasil dari *Chi Square* menghasilkan nilai signifikansi pada setiap fitur, yang mencerminkan fitur tersebut berkorelasi dataset aslinya. Hasil dari *Chi Square* gunakan sebagai dasar dalam proses *feature selection* pada SVM untuk menganalisis sentimen *tweet*. *Feature selection* bertujuan memilih subset fitur yang paling relevan dari data. Dengan menggunakan nilai signifikansi dari *Chi Square*,

SVM berfokus pada fitur yang memiliki dampak signifikan pada sentimen tweet. Akhirnya, penelitian menganalisis akurasi *Chi Square* dalam penentuan *feature selection* pada SVM. Akurasi dapat diukur berdasarkan *accuracy score, precision score, recall score* dan *f1 score*.