BAB II

TINJAUAN PUSTAKA DAN DASAR TEORI

2.1 Tinjauan Pustaka

Penelitian yang dilakukan Hendrastuty dkk. (2021), penelitian ini untuk membandingkan keakuratan dari dua fungsi pemisah (kernel) yaitu linear dan RBF dalam melakukan identifikasi opini publik pada program Kartu Prakerja, dari 2.000 data yang berhasil di kumpulkan menggunakan metode scraping pada platform X diberi label sentimen oleh ahli bahasa (*manual*), hasil pengujian kemudian dievaluasi dengan *confusion matrix* dengan ordo 3x3 maka hasil *kernel* linear mencapai akurasi 98.67% dan RBF 98.34%, penelitian ini menunjukkan jika SVM sangat efektif digunakan untuk klasifikasi teks dan *kernel* linear menjadi pemisah lebih baik dibandingkan *RBF*.

Penelitian yang dilakukan Damayanti dkk. (2024) di mana SVM digunakan untuk melakukan klasifikasi dan analisis opini publik terkait kebijakan pemerintah pada BPJS. Dengan data sebanyak 3.060 yang sudah berlabel diperoleh dari situs Kaggle, *preprocessing* dilakukan dan kemudian *undersampling* dan *oversampling* dilakukan agar model tidak bias akibat sebaran data yang tidak merata. Penelitian ini menghasilkan akurasi SVM mencapai 94,28%, F1-score dari kedua kelas masih menunjukkan perbedaan yang cukup signifikan di mana kelas positif hanya 39% dan negatif mencapai 97%.

Penelitian Tukino dan Fifi (2024) yang bertujuan melakukan klasifikasi opini masyarakat di Kota Batam terkait layanan Gojek di kota tersebut. Penelitian ini menggunakan *5-fold cross-validation* untuk menghasilkan model yang dapat melakukan klasifikasi dengan objektif dan mengurangi bias dalam proses pengujian. Hasilnya SVM mencapai akurasi 88,5% dan model dapat dengan baik mengenali teks dengan sentiment positif dengan baik di mana F1-score tertinggi berada di kelas positif, 89,5%.

Dalam beberapa penelitian juga menunjukkan bahwa *Support Vector Machine* lebih unggul dalam tugas klasifikasi teks, penelitian yang dilakukan

Septiana dan Alita (2024) dengan membandingkan SVM dengan *Random Forest* untuk melakukan klasifikasi opini publik terkait perhitungan cepat pemilu tahun 2024, data sebanyak 2.000 di bersihkan lalu diberi label secara otomatis dengan pendekatan *lexicon* dan setiap kata diberi bobot dengan *count vectorizer*. Hasil penelitian ini menunjukkan bahwa SVM lebih unggul dengan akurasi 80% dibandingkan dengan *Random Forest* yang hanya mencapai akurasi 78%. *Support Vector Machine* juga terbukti dalam penelitian Pamungkas dan Kharisudin (2021) yang melakukan klasifikasi opini publik terkait pandemi covid-19 dengan menggunakan tiga algoritma *machine learning* yaitu *Support Vector Machine, Naïve Bayes, K-Nearest Neighbor* atau KNN. Dengan pendekatan berbasis lexicon dan klasifikasi menggunakan tiga algoritma tersebut yang dievaluasi dengan *10-fold cross-validation* maka hasil menunjukkan SVM dengan *kernel* linear memperoleh rata-rata akurasi tertinggi 90,1%, dibandingkan dengan *Naïve Bayes* 79,2% dan KNN 62,1%

Berdasarkan penelitian terdahulu yang telah dipaparkan, dapat diketahui jika algoritma *Support Vector Machine* memiliki performa yang konsisten dan baik dalam melakukan klasifikasi opini publik pada berbagai permasalahan. Namun terdapat beberapa perbedaan pada penelitian ini, pada penelitian pelabelan dilakukan secara otomatis berbasis *lexicon* (lexicon based), berbeda dengan yang dilakukan Hendrastuty dkk (2021) yang menggunakan metode pelabelan manual dengan melibatkan ahli bahasa. Perbedaan selanjutnya yaitu TF-IDF digunakan untuk memberikan representasi angka (numerik) dari setiap teks, berbeda dengan yang dilakukan Septiana dan Alita (2024) yang menggunakan metode *Count Vectorizer* untuk ekstraksi fitur. Perbedaan terakhir terletak pada skema evaluasi model di mana penelitian ini menggunakan *hold-out validation* dengan membagi data menjadi dua bagian dengan rasio tertentu, skema ini berbeda dengan yang dilakukan Tukino dan Fifi (2024) dan Pamungkas dan Kharisudin (2021) yang menggunakan skema *k-fold cross-validation* untuk melakukan evaluasi model. Ringkasan dari tinjauan penelitian terdahulu sebagai dasar acuan dalam penelitian ini dapat dilihat pada Tabel 2.1.

Tabel 2.1 Penelitian terdahulu

No	Nama	Topik	Metode	Hasil
1	(Hendrastuty dkk. 2021)	Identifikasi opini publik terhadap program kartu pra kerja	SVM, *kernel linear* dan RBF	*Kernel linear* 98.67%, RBF 98.34%
2	(Damayanti dkk. 2024)	Klasifikasi opini publik terhadap program BPJS	SVM, *undersampling* dan *oversampling*	Akurasi model mencapai 97%, namun meskipun sudah dilakukan penanganan ketidakseimbangan data, nilai F1-score pada kedua kelas masih berbeda jauh (positif 39% dan negatif 97%).
3	(Tukino dan Fifi, 2024)	Melakukan klasifikasi dan analisis terhadap layanan Gojek di kota Bata	SVM, *k-fold cros- validation*	akurasi 88,5%, F1-score tertinggi di kelas positif 89,5%.
4	(Septiana & Alita, 2024)	Klasifikasi opini publik terhadap hasil perhitungan cepat	SVM, *Random Forest***,** *count vectorizer, lexicon based*	SVM mencapai hasil tertinggi 80%, Random Forest hanya mencapai 78%.
5	(Pamungkas & Kharisudin, 2021)	Klasifikasi opini publik terhadap isu covid-19	*SVM, Naïve Bayes, KNN, k-fold cross-validation*	SVM mencapai hasil tertinggi 90,1%, dibandingkan dengan Naïve Bayes 79,2% dan KNN 62,1%
6	(Petra Aldevand Hosyo, 2025)	Klasifikasi opini publik terhadap peluncuran Danantara	Lexicon-based, SMOTE (penyeimbangan) , (pelabelan), TF-IDF (ekstraksi fitur), SVM kernel linear	Akurasi 81,86%, F1-score tertinggi di kelas positif 87%, kelas negatif 68%, mayoritas opini publik cenderung positif

2.2 Dasar Teori

Bagian ini menyajikan landasan teoritis yang krusial untuk perancangan, implementasi, dan interpretasi penelitian analisis sentimen. Pembahasan mencakup

konsep kunci dan prinsip kerja metode yang digunakan, seperti pengumpulan data media sosial, pra-pemrosesan teks, ekstraksi fitur TF-IDF, penanganan ketidakseimbangan data (SMOTE), dan klasifikasi *Support Vector Machine* (SVM). Kerangka teori ini menjadi dasar pemahaman alur kerja sistem dan interpretasi temuan ilmiah.

2.2.1 Data Mining

Menurut Rahayu dkk. (2018) Data *mining* atau penambangan data adalah sebuah proses pengolahan data yang besar dan kompleks untuk menemukan pola-pola tersembunyi yang dapat ditransformasikan menjadi sebuah informasi berharga. Rahayu. dkk (2018) menjelaskan bahwa Data *mining* bekerja dengan beberapa teknik berikut:

1. Klasifikasi: Ini merupakan teknik untuk mengkategorikan data berdasarkan label yang sudah diberikan sebelumnya.
2. Regresi: Ini adalah teknik untuk melakukan prediksi terhadap suatu hal berdasarkan hasil analisis berbagai faktor yang mempengaruhinya.
3. Klastering: Teknik ini hampir sama dengan klasifikasi, namun teknik ini adalah untuk melakukan pengelompokkan berdasarkan kemiripan karakteristik data.
4. Asosiasi: Teknik ini adalah untuk menemukan korelasi antara suatu suatu atribut dalam data.

Lebih lanjut, (Rahayu dkk., 2024) menjelaskan bahwa data mining adalah tahapan yang cukup penting dalam proses *Knowledge Discovery in Databases* (KDD), adapun tahapanya sebagai berikut:

1. Selection dan Preprocessing: Ini merupakan tahap mencari dan mengambil data-data yang relevan dan mempelajari serta memahami data sebelum proses selanjutnya, proses ini juga bertujuan menghilangkan nilai yang kosong agar data siap di proses untuk tahap selanjutnya
2. Data Cleaning dan Understanding: Ini merupakan tahap di mana data dipelajari lebih dalam dan menyeluruh untuk mengetahui karakteristik data,

proses ini menerapkan metode statistik, visualisasi data, dan pemahaman data secara mendalam untuk menangani data yang bermasalah.

3. Data Transformation: Selanjutnya data akan ditransformasikan ke dalam format atau bentuk yang lebih mudah dipahami untuk di analisis.
4. Pemodelan (Data Mining): setelah transformasi, tahap utama dalam data mining ini adalah pemodelan, menggunakan metode data mining seperti klasifikasi, klastering, ataupun asosiasi dan juga regresi untuk melakukan pengolahan terhadap data-data tersebut untuk menemukan pola-pola yang akan menjadi sumber informasi.
5. Evaluasi Hasil: Proses ini merupakan tahap dimana mengecek dan mengetahui performa dari model yang telah dibangun, proses ini umumnya menggunakan metrik akurasi.
6. Interpretation dan Visualization: Hasil pemodelan di visualisikan agar mudah untuk memahami hasil pola-pola tertentu yang dihasilkan model kemudian buat kesimpulan mengenai temuan kepada pihak atau instansi terkait tersebut.
7. Knowledge Utilization: Hasil temuan kemudian di jadikan sebagai acuan dalam pengambilan keputusan bagi pihak dan pemangku kebijkan terkait.

Untuk mengolah data opini publik yang besar dan tidak terstruktur dari media sosial, proses data mining sangat relevan digunakan. *Text mining* merupakan cabang ilmu dari data mana yang secara spesifik berfungsi untuk mengolah dan menggali informasi dari data yang tidak terstruktur seperti teks, pada penerapannya text mining sering digunakan untuk melakukan klasifikasi opini seseorang terhadap suatu isu, peristiwa, produk, dan permasalahan atau yang dikenal analisis sentimen (Rahayu dkk., 2024).

2.2.2 Analisis Sentimen

Analisis sentimen atau opinion mining merupakan suatu metode untuk yang digunakan untuk melakukan analisis dan klasifikasi pendapat atau opini dari sebuah teks (Purnamasari dkk., 2023). Analisis sentimen dinilai merupakan metode yang

efektif untuk mengukur tingkat keberhasilan suatu program atau layanan yang diluncuran (Ramlan dkk. 2023)

Analisis sentimen umumnya bisa dilakukan dengan tiga tingkatan atau level, pertama adalah level dokumen di mana sentimen dapat ditentukan berdasarkan pemahaman teks secara keseluruhan. Kedua adalah kalimat, di mana sentimen ditentukan berdasarkan tiap kalimat dari sebuah teks, yang terakhir adalah level aspek. Level ini yang berfokus pada identifikasi aspek atau bagian spesifik dari suatu objek. Dalam praktiknya, ada pendekatan yang digunakan sebagai cara untuk melakukan analisis sentimen, beberapa seperti:

1. Pendekatan Berbasis Machine Learning
   Metode ini memanfaatkan algoritma machine learning untuk melakukan analisis sentimen pada data yang sudah diberi label
2. Pendekatan Berbasis Lexicon
   Metode ini memanfaatkan kamus kata atau *lexicon* yang dipakai sebagai acuan dalam menentukan sebuah sentimen pada teks.
3. Pendekatan Hybrid
   Metode ini adalah gabungan elemen dari pendekatan *machine learning* dan *lexicon* untuk melakukan klasifikasi pada data teks.

Pendekatan yang digunakan dalam penelitian ini adalah *hybrid*, di mana *lexicon* digunakan untuk melabeli teks dan proses klasifikasi menggunakan algoritma *machine learning* yaitu *support vector machine*. Analisis sentimen umumnya dilakukan pada platform sosial media di mana opini publik sangat dinamis dan subjektif. Umumnya salah satu platform sosial media yang sering digunakan dalam analisis sentimen adalah X (Twitter) karena menyediakan data secara *real-time* dan opini yang disampaikan dalam sebuah teks yang singkat padat dan jelas.

2.2.3 X (Twitter)

X atau yang dahulu lebih dikenal Twitter merupakan platform sosial media yang memungkinkan penggunanya untuk saling berinteraksi seperti membuat, melihat

dan membalas unggahan atau yang biasa dikenal *tweet* (Fitriansyah & Sibaroni, 2023).

X merupakan salah satu sosial media yang sering digunakan baik praktisi atau peneliti untuk dijadikan sumber data karena jumlah penggunanya yang banyak dan juga X menjadi platform yang dapat menggambarkan opini publik bersifat subjektif terhadap suatu peristiwa atau isu secara *real-time*, X juga memiliki format teks yang singkat dan padat dikarenakan platform ini memiliki batas 280 karakter dalam sekali tweet ini akan menuntut penggunanya untuk menyampaikan gagasan secara singkat padat dan langsung pada intinya atau *straight to the point* (Purnamasari dkk. 2023).

Sumber data teks yang berasal dari sosial media seperti X akan banyak memiliki bentuk yang tidak formal dan tidak terstruktur maka diperlukan suatu teknologi yang dapat memproses data teks ini agar dapat dipahami dan diolah oleh komputer. *Natural Language Processing* atau yang lebih dikenal NLP merupakan teknologi yang dapat memproses bahasa manusia agar dapat dipahami oleh komputer.

2.2.4 Natural Language Processing (NLP)

Natural Language Processing adalah sebuah subbagian dari *Artificial Intelligence* yang memungkinan komputer untuk dapat memahami bahasa manusia dan juga membuat komputer dalam memberikan pemahaman dengan bahasa yang dapat dimengerti oleh manusia, sederhananya NLP menjadi jembatan antara manusia dan komputer untuk saling berinteraksi (Nurwanda dkk. 2024). Pra-pemrosesan Teks merupakan proses untuk membersihkan data berupa teks sebelum di analisis lebih lanjut, terdapat beberapa tahapan dalam proses ini seperti:

1. Case Folding: Mengubah format seluruh teks menjadi huruf kecil.
2. Cleaning: Membersihkan elemen-elemen yang tidak penting seperti URL, *mention*, emoji, angka,
3. Cleansing: Hapus seluruh tanda baca seperti titik, koma, tanda tanya dan seterusnya.
4. Stopword Removal: Hapus kata-kata penting yang tidak memiliki arti dalam bahasa Indonesia seperti dan, yang, di, dan seterusnya.

5. Stemming: Setiap kata bahasa Indonesia dirubah ke kata dasar seperti berlari, pelari, pelarian akan diubah menjadi lari dan juga kata seperti kebijakan, berkebijakan akan diubah menjadi bijak.
6. Joining: Mengembalikan data teks menjadi bentuk *string*.
7. Filtering: Menghapus baris data yang kosong setelah tahap pra-pemrosesan teks.

Setelah data teks dibersihkan, agar model algoritma yang dipakai dapat mengolah teks tersebut maka teks harus diubah ke dalam bentuk angka (numerik), TF IDF merupakan salah satu metode yang bisa digunakan untuk mengubah teks menjadi format angka.

2.2.5 Ekstraksi Fitur

Ekstraksi fitur merupakan tahap di mana data teks diubah menjadi format yang dapat dipahami oleh algoritma *Support Vector Machine*, metode ekstraksi fitur yang dipakai adalah TF IDF karena efektivitasnya dalam merepresentasikan fitur berbasis kata-kata dan juga pendekatannya yang fokus melakukan ekstraksi fitur pada tingkat kata (Nurwanda dkk. 2024).

Term frequency Inverse Document Frequency atau TF IDF merupakan sebuah metode statistik yang diterapkan dalam pra-pemrosesan teks dengan mengubah setiap kata menjadi bentuk angka (numerik) agar dapat dipahami oleh algoritma SVM (Nurwanda dkk. 2024), TF IDF bekerja dengan cara mengubah setiap kata menjadi bentuk angka untuk menentukan seberapa penting sebuah kata dalam dokumen dengan seluruh kumpulan dokumen. TF IDF terdiri dari dua yaitu:

1. Term frequency (TF): Untuk menghitung seberapa kata sering muncul dalam sebuah dokumen.
2. Inverse Document Frequency (IDF): Mencari dan mengurangi kata-kata yang sering muncul dibanyak dokumen maka dianggap kurang informatif.

TF IDF dapat dihitung dengan persamaan:

$$TF - IDF(t,d) = TF(t,d) \times \log\left(\frac{N}{DF\ (t)}\right) \quad (1)$$

t: kata tertentu.

d: adalah dokumen tertentu.

N: jumlah total dokumen dalam korpus.

$DF(t)$: jumlah dokumen yang mengandung kata t.

Term Frequency Inververse Document Frequany (TF-IDF) dipilih sebagai teknik dalam ekstraksi fitur didasari oleh beberapa alasan berikut:

1. Mereduksi kata yang tidak relevan

   Meskipun dalam *StopWord Removal* telah dilakukan, namun beberapa kata penting yang merujuk pada konteks atau topik dalam penelitian akan berpotensi muncul di banyak dokumen dan akan dianggap kata-kata umum, contohnya "investasi", "ekonomi", "keuangan", oleh sebab itu TF-IDF digunakan untuk mengurangi kata-kata umum tersebut dan fokus pada kata yang penting (sering muncul di sebuah dokumen tapi jarang muncul di dokumen lain di seluruh korpus).

2. Efektif pada konteks klasifikasi biner SVM

   Data yang dihasilkan proses TF-IDF umumnya bersifat linear (linearly separable) yang akan cocok dan dengan konsep klasifikasi yang dilakukan SVM menggunakan *kernel* linear (Joachims, 2018).

Setelah teks diubah menjadi bentuk angka (numerik) maka data bisa dikatakan siap diolah oleh algoritma model yang dipilih. *Support Vector Machine* dipilih karena dinilai merupakan algoritma *machine learning* yang paling efektif untuk klasifikasi teks.

2.2.6 Machine Learning

Machine Learning adalah subbidang dari *Artificial Intelligence* bertujuan untuk mengajari atau memprogram sebuah sistem dapat belajar dari data yang diberikan lalu melakukan tugas seperti prediksi dan analisis secara mandiri tanpa harus di jalankan oleh secara manual oleh manusia (Cholissodin dan Soebroto, 2021). Machine Learning memiliki beberapa metode dalam mempelajari data yang diberikan, metodenya meliputi:

1. Supervised Leaning: ini merupakan metode di mana sistem atau model diberi data berlabel untuk dipelajari pola-pola di dalamnya, hasilnya

pembelajaran tersebut dipakai sebagai acuan bagi model untuk prediksi atau analisis pada data selanjutnya. Contohnya untuk sistem pendeteksian email spam, sistem akan diberi data berlabel berupa contoh email spam dan bukan spam agar sistem dapat belajar dari contoh tersebut untuk kemudian diterpakan ke data selanjutnya. Umumnya algoritma yang digunakan seperti Support Vector Machine, Naïve Bayes, Random Forest.

2. Unsupervised Learning: Adalah Model yang belajar untuk menemukan pola secara mandiri tanpa harus diberi data berlabel sebelumnya. Contohnya untuk menentukan tren segmentasi pasar, sistem diberi data mentah lalu dipelajari dengan mencari kesamaan pada pola-pola tertentu lalu dikategorikan atau yang biasa disebut *Clustering*. Umumnya algoritma yang digunakan adalah K-Means Clustering, Principal Component Analysis (PCA), Hierarchical Clustering.

2.2.7 Pembagian Data

Pembagian data dalam machine learning adalah tahap di mana data dibagi menjadi dua bagian yaitu dataset pelatihan dan pengujian dengan rasio tertentu yang bertujuan agar sistem atau model yang dibangun tidak memiliki bias dalam saat dilatih dan agar model dapat memahami dan memprediksi dan analisis pada data baru dengan baik (Muraina, 2022). Lebih lanjut Muraina (2022) menuturkan bahwa salah satu rasio pembagian data yang umumnya digunakan adalah 80:20 karena dapat memberikan keseimbangan yang baik antara proses pelatihan dan pengujian.

2.2.8 Penanganan Data Tidak Seimbang

Untuk mendapatkan performa yang baik dalam model klasifikasi, keseimbangan pada data diperlukan agar model dapat melakukan proses klasifikasi secara objektif. Namun ketika data mengalami kesenjangan yang cukup signifikan pada kelas mayoritas dan minoritas, hal ini akan menyebabkan model cenderung bekerja lebih baik pada kelas mayoritas dan mengabaikan kelas minoritas serta menghasilkan akurasi yang tidak valid (Fiddin, 2025).

Menurut Fiddin (2025) Chawla menjelaskan bahwa *Synthetic Minority Over-sampling Technique* atau yang dikenal SMOTE adalah teknik penanganan data yang tidak seimbang dengan membuat sampel data tiruan yang mirip dari kelas minoritas untuk menyeimbangkan sebaran data dalam kedua kelas.

Dalam penelitian ini SMOTE hanya akan diterapkan pada data latih untuk menjaga data uji tetap asli dan merepresentasikan data baru yang belum pernah diketahui oleh model sebelumnya. Jika SMOTE diterapkan terlebih dahulu pada seluruh sebelum pembagian data maka data uji dapat terpengaruh oleh sampel sintetis SMOTE dan akan berdampak pada akurasi model yang terlihat tinggi namun tidak valid karena model sudah mengetahui sebagian isi data uji sebelumnya (Ridwan dkk. 2024).

2.2.9 Support Vector Machine

Support Vector Machine adalah sebuah algoritma *machine learning* berbasis *supervised learning* di mana model belajar dari data yang sudah diberi label kemudian hasilnya dipakai sebagai acuan dan referensi dalam melakukan pengujian pada data baru. (Cholissodin dan Soebroto, 2021).

Pada dasarnya SVM dibangun untuk mengatasi masalah klasifikasi *liniear*, namun dengan adanya fungsi *kernel* membuat SVM juga bisa dipakai untuk tugas klasifikasi yang sifatnya non-linear (Cholissodin dan Soebroto, 2021). *Support Vector Machine* adalah sebuah algoritma *supervised learning* yang belajar dari pola yang dikasi untuk selanjutnya dijadikan referensi untuk melakukan pengujian pada data baru.

Dalam konteks klasifikasi linear yang divisualisasikan pada Gambar 2.1, SVM bekerja dengan cara mencari garis pemisah (*hyperplane*) terbaik, *hyperplane* terbaik adalah yang dapat memisahkan kedua bidang kelas data dengan jarak yang maksimal (margin maksimal) yang didapati dengan cara menghitung jarak antara *hyperplane* dengan titik data terdekat dari kedua kelas (*support vector*). Cholissodin dan Soebroto (2021) menjelaskan bahwa untuk mencari *hyperplane* dalam konteks klasifikasi linear, maka dapat ditulis dengan persamaan sebagai berikut init:

$$f(x) = w \cdot x + b \quad (2)$$

Keterangan:

w: adalah bobot vector

b: adalah bias

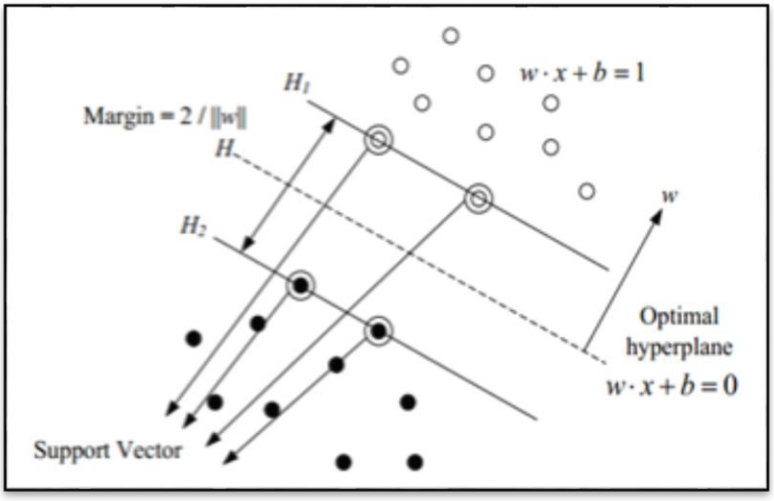
Support Vector Machine memiliki tujuan untuk mencari nilai w dan b yang paling baik untuk membuat garis pemisah atau hyperplane yang dapat memisahkan kedua kelas data dengan baik atau margin maksimal. Kedua nilai tersebut diperoleh dengan cara memperkecil fungsi objektif:

$$\frac{1}{2}\|W\|^2 \quad (3)$$

untuk mendapatkan jarak yang jauh atau margin maksimal. Namun setiap data harus berada dalam kelas sesuai labelnya agar memenuhi syarat:

$$y_i(w \cdot x_i + b) \geq 1 \quad (4)$$

agar pemisah (hyperplane) dapat bekerja dengan optimal dalam melakukan klasifikasi pada kedua kelas (Cholissodin & Soebroto, 2021).



Gambar 2.1 SVM linear (Cholissodin & Soebroto, 2021)

Dalam praktiknya tidak semua data dapat dipisahkan secara linear, oleh karena itu SVM memiliki metode untuk mengatasi hal ini dengan fungsi *kernel*, metode ini memungkinkan agar data diubah ke dalam dimensi yang lebih tinggi, ada banyak jenis *kernel* berikut ini merupakan beberapa jenis *kernel* yaitu:

1. Kernel Linear:
$$K(x, y) = x^T y \tag{5}$$
2. Kernel Polynomial:
$$K(x, y) = (x^T y + C)^d \tag{6}$$
3. Kernel RBF (Radial BasisFunction):
$$K(x, y) = \exp\left(-\frac{||x - y||^2}{2\sigma^2}\right) \tag{7}$$

Pemilihan algoritma *Support Vector Machine* dan *kernel* linear dikarenakan SVM pada konsepnya dasarnya adalah algoritma klasifikasi linear (Cholissodin dan Soebroto, 2021) dan juga hasil ekstraksi fitur teks yang dilakukan dengan TF IDF umumnya bersifat *linearly separable* maka *kernel* linear sudah sangat efektif digunakan untuk melakukan klasifikasi teks (Joachims, 2018)

2.2.10 Pelabelan

Dalam penelitian ini proses klasifikasi banyak hanya difokuskan pada dua polaritas yakni positif dan negatif atau klasifikasi biner. Pendekatan ini didasari oleh penelitian yang dilakukan oleh Pang dkk. (2002), mereka juga menyampaikan bahwa dasar dalam analisis sentimen adalah untuk melakukan klasifikasi antara opini positif dan negatif.

Penelitian ini menggunakan metode berbasis *lexicon* untuk melakukan pelabelan sentimen, metode *lexicon* adalah salah satu metode yang dapat dipakai dalam bidang penambangan data untuk melakukan klasifikasi teks berdasarkan kamus yang berisi daftar kata berkonotasi positif, negatif, dan netral (Ratnaswari dkk., 2025). Kamus *lexicon* yang dipakai dalam penelitian adalah Indonesia Sentimen *Lexicon* (InSet Lexicon) yang dibuat oleh Koto dan Rahmaningtyas (2017) guna mendukung tugas klasifikasi teks berbahasa Indonesia yang berbasis data teks

dengan format ringkas, dan singkat. Pemilihan kamus *InSet* untuk digunakan dalam penelitian ini didasari beberapa keunggulannya sebagai berikut:

1. Relevansi

   Indonesia Sentiment Lexicon (InSet Lexicon) dibangun bukan dari hasil terjemahan, kamus ini merupakan kamus sentimen yang dibangun dari teks berbahasa Indonesia yang diperoleh dari platform X, ini artinya kata dan polaritasnya akan lebih relevan jika diterapkan dalam tugas klasifikasi teks yang berbahasa Indonesia (Koto & Rahmaningtyas, 2017).

2. Pemberian bobot polaritas

   Kamus *InSet* tidak hanya berisi daftar kata dan label polaritasnya, namun juga memiliki bobot polaritas pada tiap labelnya untuk semakin memperjelas seberapa kuat polaritas suatu kata pada label tersebut (Koto dan Rahmaningtyas, 2017).

Namun selain kelebihan yang telah disampaikan, Koto dan Rahmaningtyas (2017) juga menjelaskan jika *InSet* memiliki beberapa kelemahan seperti:

1. Kecenderungan bias negatif

   Inonesian Sentiment Lexicon memiliki kelemahan yang disebabkan perbedaan jumlah kata yang terdapat di dalamnya, di mana daftar kata positif berjumlah 3.609 dan negatif 6.607. Perbedaan ini bisa menimbulkan bias pada sentimen negatif.

2. Sensitif terhadap sinonim kata

   Kelemahan selanjutnya adalah *InSet* sangat sensitif terhadap sinonim kata. Meskipun dalam proses pembuatannya, Koto dan Rahmaningtyas (2017) telah menambahkan sinonim dari setiap kata positif maupun negatif yang telah berada di kamus namun, terdapat beberapa sinonim yang tidak konsisten dengan sentimennya. Artinya, dua kata yang dianggap memiliki makna dasar yang sama (sinonim) tetapi berpotensi memiliki bobot polaritas atau sentimen yang berbeda. Sebagai contoh kata “bicara” dan “bacot” secara makna dasarnya sama tetapi bobot polaritasnya bisa berbeda.

Pada penerapannya klasifikasi biner tersebut pelabelan dilakukan secara otomatis menggunakan metode berbasis atau leksikon *based*. Metode pelabelan penelitian ini diadopsi dari penelitian Septiana dan Alita (2024) yang memberi label berdasarkan selisih skor antara kata positif dan negatif yang ditemukan dalam teks.

1. Perhitungan Skor:

Perhitungan skor dilakukan dengan menjumlahkan bobot polaritas setiap teks dalam data (t) yang ditemukan dalam kamus positif (P) dan kamus negatif (N) dan total skor untuk setiap teks $(total_{score_t})$ didapatkan dari hasil perhitungan bobot teks yang terdapat dalam kamus positif dan negatif. Jika sebuah kata dalam teks w ditemukan di kamus positif maka skornya $score(w)_P$ namun jika sebuah kata dalam teks ditemukan dalam kamus negatif maka skornya adalah $score(w)_N$, perhitungan skor dapat dilakukan dengan persamaan berikut ini:

$$total_{score_t} = \sum_{w \in t \cap P} score(w)_P + \sum_{w \in t \cap N} score(w)_N \quad (8)$$

Keterangan:

$total_{score_t}$: adalah total skor sentimen untuk teks (*tweet*) t.

w: adalah kata yang terdapat dalam teks t.

P: adalah kamus kata positif.

N: adalah kamus kata negatif.

$score(w)_P$: adalah jumlah bobot polaritas kata w jika terdapat dalam kamus positif P.

$score(w)_N$: adalah jumlah bobot polaritas kata w jika terdapat dalam kamus negatif N.

2. Pemberian label

Setelah skor total sentimen didapatkan, maka teks akan dilabel dengan aturan logika berikut:

a. Positif: jika $total_{score_t} > 0$.
b. Negatif: jika $total_{score_t} < 0$.
c. Netral: jika $total_{score_t} = 0$.

2.2.11 Evaluasi Hasil

Evaluasi hasil dilakukan untuk mengukur performa model yang dibangun apakah sudah sesuai dengan keinginan solusi yang dicapai atau tidak, untuk melakukan evaluasi pada model algoritma *machine learning* yang bangun diperlukan beberapa metrik atau penilaian yang harus seperti *Precision, recall, F-measure, Sensitivity, Spesificity,* dan *Accuracy* (Purnamasari dkk. 2023).

1. Accuracy: Ini adalah hasil perhitungan dari semua jenis nilai yang didapatkan.

$$Accuracy = \frac{TP + TN}{\dot{T}P + FP + \bar{F}P + TN} x100\% \quad (9)$$

2. Precision: Ini adalah seberapa akurat model dapat mengenali dan menebak seluruh teks di kelas positif.

$$Precision = \frac{TP}{TP + FP} x100\% \quad (10)$$

3. Recall: Adalah sebarapa model bisa menangkap semua teks di kelas positif yang di temukan.

$$Recall = \frac{TP}{TP + FN} \quad (11)$$

4. F1- Score: Adalah keseimbangan antara precision dan recall.

$$F - Measure = \frac{2 * Precision * Recall}{Precision + Recall} \quad (12)$$

Kelima metrik untuk evaluasi tersebut didapatkan dari hasil ukuran utama yang dihasilkan oleh confusion matrix seperti berikut ini:

1. True Positives (TP): Jumlah prediksi yang benar pada kelas positif.
2. True Negatives (TN): Jumlah prediksi yang benar pada kelas negatif.
3. False Positives (FP): Jumlah prediksi yang salah pada kelas positif.
4. False Negatives (FN): Jumlah prediksi yang salah dikelas negatif.