BAB II
LANDASAN TEORI

2.1 Tinjauan Pustaka

Dalam pembuatan penelitian ini ada penelitian-penelitian yang sebelumnya sudah pernah dilakukan oleh orang lain yang mirip dan bahkan menjadi acuan dari penelitian ini. Adapun beberapa penelitian tersebut antara lain sebagai berikut :

Tabel 2. 1 Tinjauan Literatur

No	Detail Jurnal	
1	Peneltian	(Styawati, Hendrastuty, *et al.*, 2021)
	Judul	Analisis Sentimen Masyarakat Terhadap Program Kartu Prakerja Pada *Twitter* Dengan Metode *Support Vector Machine*
	Metode	*Support Vector Machine (SVM)*
	Hasil	Hasil evaluasi yang dilakukan pada nilai akurasi kernel linear 98.67%, precission 98%, recall 99%, dan F1-Score 98%, sedangkan pada nilai akurasi kernel RBF 98.34%, precission 97%, recall 98%, F1- Score 98%, dapat disimpulkan bahwa sentimen masyarakat dari pengguna twitter terhadap program kartu prakerja dimasa pandemi lebih condong ke netral sebesar 98,34%. Berdasarkan hasil evaluasi yang dilakukan pada nilai akurasi kernel linear menghasilkan nilai akurasi 98.67%, sedangkan kernel RBF menghasilkan akurasi 98.34%. Maka dari sisi akurasi kernel linear lebih akurat dari pada kernel RBF.
	Perbedaaan	Hasil evaluasi yang dilakukan pada nilai akurasi kernel linear 77,70%, Precission 69%, recall 78%, dan F1-Score 69%, sedangkan pada nilai akurasi kernel RBF 77,70%, Precission 69%, recall 78%, dan F1-Score 69%, dapat disimpulkan bahwa sentimen terhadap penggunaan media pembelajaran ai lebih condong suka pembelajaran ai sebesar 77,70%, berdasarkan hasil evaluasi yang dilakukan pada nilai akurasi kedua kernel menghasilkan nilai akurasi 77,70%.

Tabel 2. 2 Tinjauan Literatur (Lanjutan)

No	Detail Jurnal	
2	Peneltian	(Rahman Isnain *et al.*, 2021)
	Judul	Sentimen Analisis Publik Terhadap Kebijakan Lockdown Pemerintah Jakarta Menggunakan Algoritma SVM
	Metode	*Support Vector Machine (SVM)*
	Hasil	Untuk mengetahui bagaimana sentiment publik terhadap kebijakan yang akan dilakukan pemerintah mengenai kebijakan lockdown ataupun pembatasan sosial berskala besar menggunakan metode *Support Vector Machine* denga ekstraksi fitur tf-idf dengan pengujian yang nantinya akan dilihat bagaimana nilai *accuracy, precision, Recall* dan *F1-Score*. Penggunaan metode *Support Vector Machine* dan ekstraksi fitur dengan tf-idf yang membagi kelas menjadi sentiment positif 68,75% dan negative 31,25% menghasilkan nilai accuracy sebesar 74%, precision sebesar 75%, recall sebesar 92% dan F1-Score sebesar 83%.
	Perbedaaan	Untuk mengetahui bagaimana sentiment pengguna terhadap media pembelajaran ai menggunakan metode Support Vector Machine dengan ekstraksi fitur Word2Vec dengan pengujian yang nantinya akan dilihat bagaimana nilai accuracy, precision, Recall dan F1-Score.Penggunaan metode Support Vector Machine dan ekstraksi fitur Word2Vec yang membagi kelas menjadi sentiment positif 77,70% dan negative 22,30% menhasilkan nilai akurasi kernel linear 77,70%, Precission 69%, recall 78%, dan F1-Score 69%, sedangkan pada nilai akurasi kernel RBF 77,70%, Precission 69%, recall 78%, dan F1-Score 69%, dapat disimpulkan bahwa sentimen terhadap penggunaan media pembelajaran ai lebih condong suka pembelajaran ai sebesar 77,70%, berdasarkan hasil evaluasi yang dilakukan pada nilai akurasi kedua kernel menghasilkan nilai akurasi 77,70%.
3	Peneltian	(Darwis, Pratiwi and Pasaribu, 2020)
	Judul	Penerapan Algoritma SVM Untuk Analis Sentiment Pada Data Twitter Komisi Pemberantasan Korupsi Republik Indonesia menggunakan metode Support Vector Machine.

Tabel 2. 3 Tinjauan Literatur (Lanjutan)

No	Detail Jurnal	
	Metode	*Support Vector Machine (SVM)*
	Hasil	Hasil dari klasifikasi menggunakan metode SVM dibagi menjadi tiga kelas, yaitu kelas positif sebanyak 8%, kelas negatif sebanyak 77%, dan kelas netral sebanyak 15%. Pengujian dari penelitian ini menggunakan Confusion Matrix, Berdasarkan hasil pengujian dan evaluasi yang dilakukan pada nilai akurasi, precession, recall, dan F1-Score, dapat disimpulkan bahwa sentimen masyarakat pengguna *twitter* mengenai kinerja KPK dengan presentase sangat kurang baik yaitu ditunjukan dengan munculnya akurasi dari hasil penelitiannya kecondongan sentimen negatif sebesar 77% dengan keakuratan hasil pengujian akurasi sebesar 82% dan pengujian precision sebesar 90%, serta recall sebesar 88% dan f1-score sebesar 89%.
	Perbedaaan	Hasil dari klasifikasi menggunakan metode SVM dibagi menjadi dua sentiment positif 77,70% dan negative 22,30% menhasilkan nilai akurasi kernel linear 77,70%, Precission 69%, recall 78%, dan F1-Score 69%, sedangkan pada nilai akurasi kernel RBF 77,70%, Precission 69%, recall 78%, dan F1-Score 69%, dapat disimpulkan bahwa sentimen terhadap penggunaan media pembelajaran ai lebih condong suka pembelajaran ai sebesar 77,70%, berdasarkan hasil evaluasi yang dilakukan pada nilai akurasi kedua kernel menghasilkan nilai akurasi 77,70%.
4	Peneltian	(Styawati, Nurkholis, *et al.*, 2021)
	Judul	Optimasi Parameter Support Vector Machine Berbasis Algoritma Firefly Pada Data Opini Film
	Metode	*Support Vector Machine (SVM)*

Tabel 2. 4 Tinjauan Literatur (Lanjutan)

No	Detail Jurnal	
	Hasil	Keberhasilan klasifikasi metode SVM bergantung pada koefisien soft margin C, serta parameter dari fungsi kernel. Parameter SVM tersebut biasanya didapatkan dengan cara trial and error, namun cara tersebut membutuhkan waktu yang cukup lama karena harus mencoba setiap kombinasi parameter SVM, maka dari itu tujuan dari penelitian ini adalah mencari nilai parameter SVM yang optimal berdasarkan akurasi. Penelitian ini menggunakan Firefly Algorithm (FA) sebagai metode optimasi parameter SVM. Data set yang digunakan dalam penelitian ini adalah data opini masyarakat terhadap beberapa film. Label kelas yang digunakan dalam klasifikasi data yaitu label kelas positif dan label kelas negatif. Banyaknya data yang digunakan dalam penelitian ini yaitu 2179 data, dengan pembagian data sebanyak 436 sebagai data pengujian dan 1743 data sebagai data pelatihan. Berdasarkan data tersebut dilakukan proses evaluasi terhadap Firefly Algorithm-Support Vector Machine (FASVM). Hasil penelitian ini menunjukkan bahwa Algoritma Firefly mampu mendapatkan kombinasi parameter SVM yang optimal berdasarkan akurasi, sehingga tidak diperlukan cara trial and error untuk mendapatkan nilai tersebut. Hal ini dibuktikan dengan hasil evaluasi FA-SVM menggunakan rentang nilai C=1.0-3.0 dan =0.1-1.0 menghasilkan akurasi tertinggi yaitu 87.84%. Evaluasi berikutnya menggunakan rentang nilai C=1.0-3.0 dan =1.0-2.0 menghasilkan akurasi tertinggi 87.15%.
	Perbedaaan	Keberhasilan klasifikasi metode SVM bergantung pada koefisien soft margin C, serta parameter dari fungsi kernel. Parameter SVM tersebut biasanya didapatkan dengan cara trial and error, namun cara tersebut membutuhkan waktu yang cukup lama karena harus mencoba setiap kombinasi parameter SVM, maka dari itu tujuan dari penelitian ini adalah mencari nilai parameter SVM yang optimal berdasarkan akurasi.Data set yang digunakan dalam penelitian ini adalah data terhadap penggunaan media pembelajaran ai. Label kelas yang digunakan dalam klasifikasi data yaitu label kelas positif dan label kelas negatif. Banyaknya data yang digunakan dalam penelitian ini yaitu 5045 data, dengan pembagian data sebanyak 1009 sebagai data pengujian dan 4035 data sebagai data pelatihan.

Tabel 2. 5 Tinjauan Literatur (Lanjutan)

No	Detail Jurnal	
5	Peneltian	(Styawati. *et al.*, 2021)
	Judul	Sentiment Analysis on Online Transportation Reviews Using Word2Vec Text Embedding Model Feature Extraction and Support Vector Machine (SVM) Algorithm
	Metode	Word2vec support vector machine (SVM)
	Hasil	Penelitian ini menggunakan model penyisipan teks word2vec dan algoritma support vector machine (SVM). Word2vec digunakan sebagai model ekstraksi ciri sebagai representasi kata ke dalam bentuk vektor. Arsitektur model word2vec yang digunakan adalah model skip-gram. Algoritma Support Vector Machine (SVM) digunakan untuk proses klasifikasi data untuk menentukan tingkat akurasi dari sentimen data yang digunakan. Hasil pengujian yang dilakukan klasifikasi analisis sentimen pada aplikasi transportasi online menunjukkan hasil kinerja yang cukup baik yaitu, aplikasi Gojek mendapatkan nilai kinerja yang lebih tinggi dengan nilai akurasi sebesar 89%, presisi sebesar 94%, re-call sebesar 86% dan f1-score sebesar 90%. Sementara itu, aplikasi Grab memiliki nilai akurasi 87%, presisi 94%, recall 85%, dan f1-score 89%. Dapat disimpulkan bahwa hasil penelitian ini menunjukkan bahwa SVM dengan kernel RBF menghasilkan akurasi tertinggi dibandingkan dengan kernel Linear, kernel Polynomial, dan kernel Sigmoid. Penerapan penyisipan teks word2vec sebagai ekstraksi fitur menggunakan model skip-gram dan algoritma SVM menggunakan kernel RBF sebagai klasifikasi.
	Perbedaaan	Penelitian ini menggunakan model penyisipan teks word2vec dan algoritma support vector machine (SVM). Word2vec digunakan sebagai model ekstraksi ciri sebagai representasi kata ke dalam bentuk vektor. Arsitektur model word2vec yang digunakan adalah model skip-gram. Algoritma Support Vector Machine (SVM) digunakan untuk proses klasifikasi data untuk menentukan tingkat akurasi dari sentimen data yang digunakan. Hasil pengujian yang dilakukan klasifikasi analisis sentimen terhadap penggunaan media pembelajaran ai yaitu, Hasil dari klasifikasi menggunakan metode SVM dibagi menjadi dua sentiment positif 77,70% dan negative 22,30% menhasilkan nilai

		akurasi kernel linear 77,70%, Precission 69%, recall 78%, dan F1-Score 69%, sedangkan pada nilai akurasi kernel RBF 77,70%, Precission 69%, recall 78%, dan F1-Score 69%, dapat disimpulkan bahwa sentimen terhadap penggunaan media pembelajaran ai lebih condong suka pembelajaran ai sebesar 77,70%, berdasarkan hasil evaluasi yang dilakukan pada nilai akurasi kedua kernel menghasilkan nilai akurasi 77,70%.

Tabel 2. 6 Tinjauan Literatur (Lanjutan)

No	Detail Jurnal	
6	Peneltian	(Alita and Fernando, 2021)
	Judul	Multiclass SVM Algorithm For Sarcasm Text In Twitter
	Metode	Multiclass Support Vector Machine (SVM)
	Hasil	Penelitian dibidang text mining sekarang ini semakin marak dilakukan karena berbagai industry dan tokoh public yang ingin mendapatkan informasi terkait pendapat publik tentang produk atau penilaian individual yang didapatkan dari media social baik pendapat yang bersifat pendapat biasa maupun sarkasme. Berdasarkan hasil penelitian yang telah dilakukan dapat disimpulkan bahwa pendeteksian sarkasme pada proses analisis sentimen dapat dilakukan dengan mengunakan metode multiclass SVM adalah SVM OAO dan SVM OAA dengan hasil nilai akurasi yang memiliki nilai yang sama besarnya baik yang dilakukan secara acak maupun tidak acak dengan nilai sebesar 60,82% dilakukan secara acak dan 60,93% secara tidak acak. Kemudian untuk nilai presisi OAA memiliki nilai unggul sebesar 88,84% yang dilakukan secara acak. Dalam hal nilai recall, OAA masih menjadi metode yang lebih baik dan nilainya dilakukan secara acak sebesar 92%. Terakhir, untuk skor F1, metode SVM OAA juga memiliki nilai yang lebih tinggi yaitu 70.91% yang dilakukan secara acak.
	Perbedaaan	Pada peneletian ini tidak menggunakan *Multiclass Support Vector Machine (SVM)* dikerenakna hanya mencari dua kelas yaitu kelas negatif dan positif.

2.2 Sentimen Analisis

Sentiment analysis atau *opinion mining* mengacu pada bidang yang luas dari pengolahan bahasa alami, *komputasi linguistic* dan *texs mining* yang memiliki tujuan menganalisa pendapat, sentimen, evaluasi, sikap, penilaian dan emosi seseorang apakah pembicara atau penulis berkenan dengan suatu topik, produk, layanan, organisasi, individu, ataupun kegiatan tertentu.

Analisis sentimen adalah cara untuk memecahkan masalah opini publik, sikap, dan emosi suatu entitas. Entitas dapat berupa individu, peristiwa, atau topik. Selain itu, analisis sentimen juga dapat digunakan untuk mengetahui mayoritas seseorang memiliki opini positif atau negatif terhadap suatu topik. Analisis sentimen juga dapat digunakan untuk menggali informasi dari suatu data, informasi tersebut nantinya dapat digunakan sebagai acuan untuk perbaikan sistem selanjutnya(Styawati. *et al.*, 2021). Hasil sentimen tersebut diambil dengan menganalisa tiap kata pada sebuah kalimat baik dari pendapat, ulasan, paragraph dan sebuah topik berkaitan dengan konteks. Analisis sentimen dilakukan berdasarkan aspek masalah yang dimiliki dan akan menentukan hasil yang bernilai positif atau netral pada sebuah kalimat dan paragraph(B Tabuhan, 2021).

Tugas utama dalam *analisis sentiment* dengan mengelompokkan teks yang ada didalam sebuah kalimat atau dokumen dengan memastikan pendapat yang ditemukan di dalam kalimat atau dokumen tersebut, apakah bersifat positif, negatif atau netral(Pudjajana dan Manongga 2018). Sentimen analisis juga dapat mengungkapkan perasaan emosional sedih, bahagia atau marah(Rusdiaman dan Rosiyadi 2019). Berdasarkan penelitian yang telah dilakukan sebelumnya, pada klasifikasi sentimen terdapat 2 bentuk jenis kelas, pertama 2 kelas yaitu positif

dan negatif dan kedua 3 kelas yaitu positif, negatif dan netral. Namun jenis kelas tersebut dapat berubah sesuai dengan kebutuhan analisis(Rahmadhani, 2021).

2.3 Text Mining

Text mining suatu ilmu komputer yang mencoba untuk memecahkan krisis informasi yang berlebihan dengan menggabungkan teknik dari data mining, pembelajaran mesin, pemrosesan bahasa alami, pencarian informasi, dan pengetahuan manajemen.

Text mining atau *text analytics* adalah istilah yang mendeskripsikan sebuah teknologi yang mampu menganalisis data teks semi-terstruktur maupun tidak terstruktur, hal inilah yang membedakannya dengan data mining dimana data mining mengolah data yang sifatnya terstruktur. Pada dasarnya, *text mining* merupakan bidang interdisiplin yang mengacu pada perolehan informasi (*information retrieval*), data mining, pembelajaran mesin (*machine learning*), statistik, dan komputasi linguistic. *Text mining* merupakan teknik yang digunakan untuk menangani masalah klasifikasi, clustering, information extraction dan information retrival. *Text mining* adalah penambangan yang dilakukan oleh komputer untuk mendapatkan sesuatu yang baru, sesuatu yang tidak diketahui sebelumnya atau menemukan kembali informasi yang tersirat secara implisit, yang berasal dari informasi yang diekstrak secara otomatis dari sumber-sumber data teks yang berbeda-beda(Rusdiaman and Rosiyadi, 2019).

Konsep teks *mining* yang digunakan dalam klasifikasi yaitu dokumen tekstual dengan tujuan untuk megkasifikasi dokumen yang sesuai dengan topic pembahasan. Perbedaan antara data mining dan *text mining* terletak pada *prepocessing*, pada data mining prepocessing berfokus pada penomoran

(*indexing*) dan normalisasi data, sedangkan *text mining* berfokus pada identifikasi dan ekstraksi fitur(Rahman Isnain *et al.*, 2021).

2.4 Twitter

Twitter didirikan oleh Jack Dorsey pada bulan Maret 2006 dengan situs jejaring sosialnya diluncurkan pada bulan Juli yang dioperasikan oleh *Twitter, Inc*. *Twitter* adalah media sosial gratis dan terpopuler serta menyediakan layanan jaringan yang memungkinkan pengguna untuk berbagi pendapat melalui pesan singkat atau sering dikenal dengan *tweet*. Ulasan dari *twitter* dapat diklasifikasikan ke dalam beberapa sentimen. Seperti penelitian yang dilakukan oleh (Styawati, Nurkholis, *et al.*, 2021) yang berjudul Optimasi parameter support vector machine Berbasis *Algoritma Firefly* pada data *opini film* diklasifikasikan ke dalam dua sentimen yaitu positif dan negatif.

Tweet dapat dilihat secara publik, namun penggunanya dapat menentukan pengiriman pesan kesiapa saja dan pengguna dapat melihat *tweet* pengguna lainnya yang biasa dikenal sebagai pengikut (*followers*). Pengguna *twitter* juga dapat menulis pesan berdasarkan topik dengan menggunakan simbol # (*hastag*). Sedangkan untuk menyebutkan nama atau membalas pesan dari pengguna lain dapat menggunakan simbol @(Rahmadhani, 2021).

Fitur yang terdapat didalam *Twitter* antara lain:

1. Laman Utama (*Home*)

Pada halaman utama pengguna dapat mengetahui *tweet* yang dikirimkan oleh pengguna lain yang telah menjadi teman kita atau *following*. Halaman utama biasa disebut sebagai *timeline*. *Timeline* ini menampilkan sebuah aliran *tweet* yang telah tersusun sesuai dengan waktu *tweet* dikirim.

2. *Box Searching*

Box Searching atau kotak pencarian merupakan fitur yang disediakan untuk bisa menemukan akun orang lain. Fitur ini juga bisa anda manfaatkan untuk menemukan kira-kira *hashtag* apa yang sedang viral atau yang tranding. Bukan hanya itu saja, fitur kotak pencarian ini juga bisa digunakan untuk menemukan akun lama orang lain yang mana mungkin anda ingin melihat beberapa *tweetnya*.

3. Profil (*Profile*)

Halaman ini yang akan dilihat oleh semua pengguna *Twitter* mengenai profil atau data diri serta *tweet* yang telah sempat dibuat.

4. Pengikut (*Followers*)

Pengikut adalah pengguna lain yang ingin menjadikan kita sebagai temannya. Ketika pengguna lain sudah menjadi pengikut akun seseorang, maka *tweet* seseorang yang telah diikuti tersebut akan muncul pada halaman utama.

5. Mengikuti (*Following*)

Mengikuti kebalikan dari pengikut, mengikuti adalah akun seseorang yang sudah mengikuti akun pengguna lain agar *tweet* yang dikirim oleh pengguna yang diikuti tersebut dapat muncul pada halaman utama.

6. *Mentions*

Biasanya konten ini merupakan balasan dari percakapan agar sesame pengguna bisa langsung menandai orang yang akan diajak bicara.

7. *Favorite*

Favorite yaitu cara untuk menyimpan sebuah *tweet* yang dianggap menarik dengan memandainya sehingga *tweet* tersebut dapat dibaca lagi suatu saat dan tidak hilang oleh halaman sebelumnya.

8. Tagar *(Hashtag)*

Hashtag "#" adalah simbol yang ditulis sebelum topik tertentu, yang digunakan agar pengguna lain dapat mencari topik yang serupa yang ditulis oleh pengguna lain juga.

9. *Tweet Activity*

Dengan adanya fitur ini, Anda bisa mengetahui berapa jumlah dari impresi dan juga interaksi serta keterlibatan para pengguna. Anda juga bisa mengetahui berapa jumlah orang yang sudah mengklik terhadap gambar maupun video yang Anda upload. Berapa jumlah *retweet*, *share*, dan lain-lainnya juga bisa diketahui melalui fitur ini.

10. List Pengguna

Twitter dapat mengelompokkan satu kelompok sehingga mempermudah untuk dapat melihat secara keseluruhan para daftar nama pengguna (*username*) yang mereka ikuti (*follow*).

11. Pesan Langsung

Pesan langsung sering dikenal dengan sebutan DM. DM sejenis inbox pada *twitter* yang langsung tertuju ke pembuat *tweet* itu sendiri.

10. *Direct message* (DM)

Memungkinkan komunikasi dua arah antar pengguna Twitter dengan batasan 140 karakter. Pihak pengirim pesan hanya bisa mengirimkan *Direct Message* pada pihak yang sudah melakukan *mem-follow* dirinya. Selain itu, Dua pengguna *Twitter* yang saling *follow* dapat berkirim *Direct Message.*

11. *Replay*

Replay sebuah balasan atas suatu *tweet* yang mengarah langsung pada si pembuat *tweet* itu.

12. Topik Hangat (*Trending Topic*)

Topik yang sedang banyak dibicarakan oleh pengguna *Twitter* dalam waktu yang bersamaan. Topik ini dapat membantu penggunanya untuk dapat mengerti apa yang sedang terjadi pada dunia.

2.5 Preprocessing Data

Data preprocessing digunakan untuk mengkondisikan dataset yang tidak terstruktur agar sesuai dengan kebutuhan sehingga data siap untuk melalui tahap selanjutnya. Tahapan yang dilakukan peneliti dalam melakukan preprocessing data adalah sebagai berikut: (Styawati. *et al.*, 2021).

Adapun preprocessing dalam penelitian ini menggunakan lima teknik yaitu case folding, cleaning, stemming, tokenization, dan stopword.

1. Case folding

Case folding adalah proses mengubah kalimat data teks ke dalam seragam. *Case folding* dilakukan dengan mengubah semua karakter dalam teks menjadi huruf kecil.

2. Cleaning

Cleaning adalah kegiatan menghilangkan karakter yang tidak sesuai dengan ketentuan yang dibuat, seperti huruf atau karakter di luar alfabet a-z, tanda baca, menghapus link atau URL, hashtag, emoticon, dan angka.

3. Stemming

Stemming data adalah proses menyaring kata-kata yang mengandung kata sambung, kata ganti, kata depan, menjadi kata dasar dengan menghilangkan *prefiks*, *sufiks*, penyisipan dan konfiks kombinasi awalan dan akhiran.

4. Tokenization

Merupakan proses seleksi pemotongan kata dalam kalimat. Diberikan tanda pemisah seperti tanda koma (,), titik (.), dan tanda pemisah lainnya. *Tokenization* berfungsi untuk memecah komentar menjadi kata-kata. Proses *tokenization* dilakukan dengan melihat setiap spasi pada komentar, kemudian berdasarkan space tersebut komentar dapat dipisah.

5. Stopword

Stopword adalah proses menghapus kata-kata yang termasuk dalam daftar stopword. *Stopwords* adalah kata-kata umum yang muncul dalam jumlah besar yang memiliki fungsi tetapi tidak memiliki arti. *Stopword* merupakan proses untuk memfilter Kata-kata yang termasuk dalam contoh stopword seperti yang, atau, mereka, kenapa, jika, seperti, untuk, ke, dan.

2.6 Word2Vec

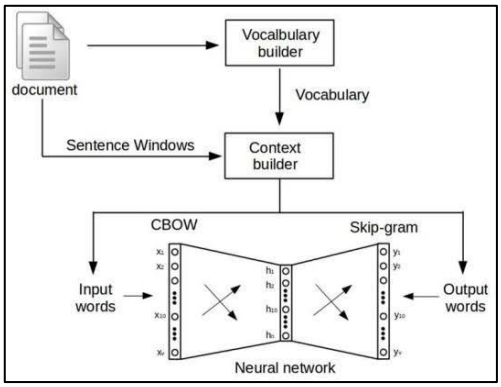
Word2vec adalah arsitektur penyisipan kata, merupakan pemetaan kata ke dalam vektor. Vektor-vektor ini nantinya akan digunakan untuk berbagai tugas *Natural Language Processing*. Pada *word2vec* dalam pembobotan kata digunakan nilai rata-rata dari vektor yang merepresentasikan kata tersebut. Model *Skip-gram Word2Vec* dapat menghasilkan akurasi yang lebih tinggi dibandingkan *Continuousbag-of word* (*CBOW*)(Styawati. *et al.*, 2021).

Word2vec dibagi menjadi dua metode yaitu *Skip gram* dan *Continuous Bag of Words (CBOW)*. *Word2vec* membuat representasi numerik terdistribusi vektor dari fitur kata. Maksud dan manfaat *word2vec* adalah mengelompokkan vektor-vektor dari kata-kata yang sejenis dalam suatu ruang vektor. Dengan menggunakan data yang memadai, *word2vec* dapat memprediksi arti kata secara akurat berdasarkan riwayat kemunculannya. Prediksi ini dapat digunakan untuk menentukan asosiasi suatu kata dengan kata lain yang mirip satu sama lain.

Model word2vec yang digunakan adalah *Skip-Gram*. *Skip Gram* memiliki kelebihan yaitu lebih cocok untuk kata-kata yang jarang muncul dibandingkan dengan CBOW. Alasan dasar pemilihan model *Skip-Gram* adalah banyaknya variasi kata, sehingga dikhawatirkan akan muncul kemunculan kata-kata yang jarang ditemukan. Penggunaan *Skip Gram* dalam masalah ini dianggap lebih tepat daripada CBOW. Ukurannya 100 dimensi dan jendelanya 5. Library yang digunakan adalah gensim.

2.7 Pembobotan Word2Vec

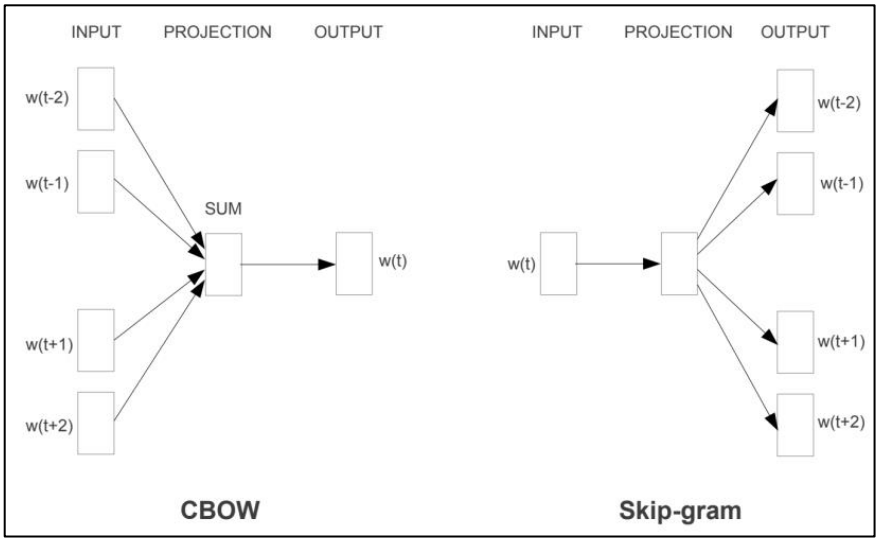
Tahapan pembobotan merupakan tahap dimana pemberian bobot pada setiap kata dengan mengunakan perhitungan Word2Vec Dalam membangun model fitur Word2vec melibatkan tiga proses yang berperan sebagai vocabulary builder, context builder, dan neural network (CBOW and Skip-gram architecture) dapat dilihat pada Gambar 3.5



Gambar 2. 1 Tahapan pembobotan Word2Vec

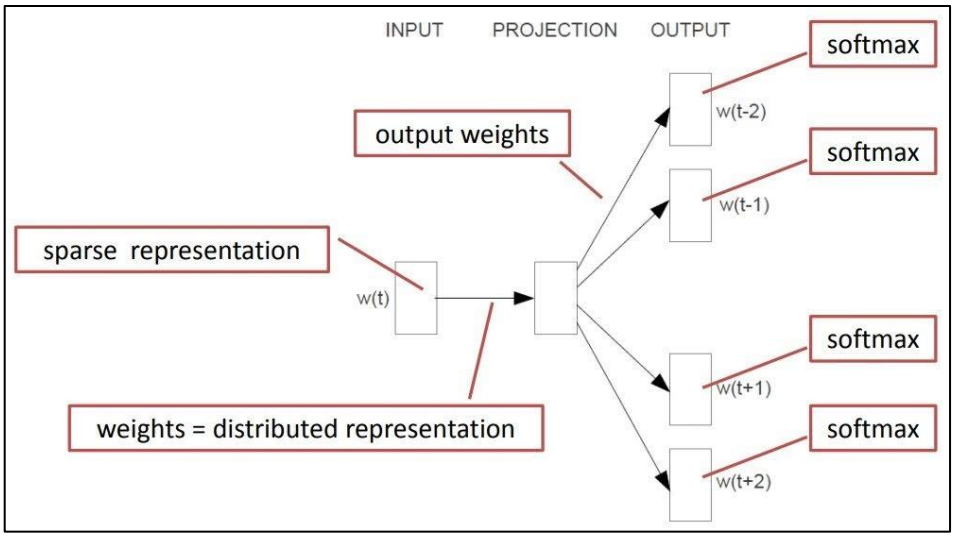
Vocabulary builder adalah blok bangunan pertama dari model word2vec. Dibutuhkan data teks mentah, sebagian besar dalam bentuk kalimat. Vocabulary builder digunakan untuk membangun kosakata dari korpus teks. Hal ini akan mengumpulkan semua kata-kata unik dari korpus dan membangun kosakata. Hasil dari proses pembangun kosakata adalah kamus kata-kata dengan indeks kata dan nilai kemunculan setiap kata.

Context builder menggunakan output dari vocabulary builder. Context builder adalah proses untuk mengetahui hubungan antara kemunculan satu kata dengan kata-kata lain di sekitarnya dengan menggunakan konsep konteks window atau disebut juga sliding window. Secara umum, ukuran jendela konteks di NLP adalah 5 hingga 8 kata tetangga. Jika kita memilih ukuran konten window adalah 5, maka akan muncul 5 kata di sebelah kiri dan kanan dari kata tengah.



Gambar 2. 2 Model Word2Vec

Skip-gram merupakan arsitektur yang menggunakan current word berperan sebagai input untuk memprediksi konteks yang berperan sebagai target terhadap sekitarnya. Skip-gram dapat mempelajari distribusi probabilitas dari kata-kata dalam sebuah konteks dengan windows yang sudah ditentukan dapat dilihat pada



Gambar 2. 3 Arsitektur Skip-gram

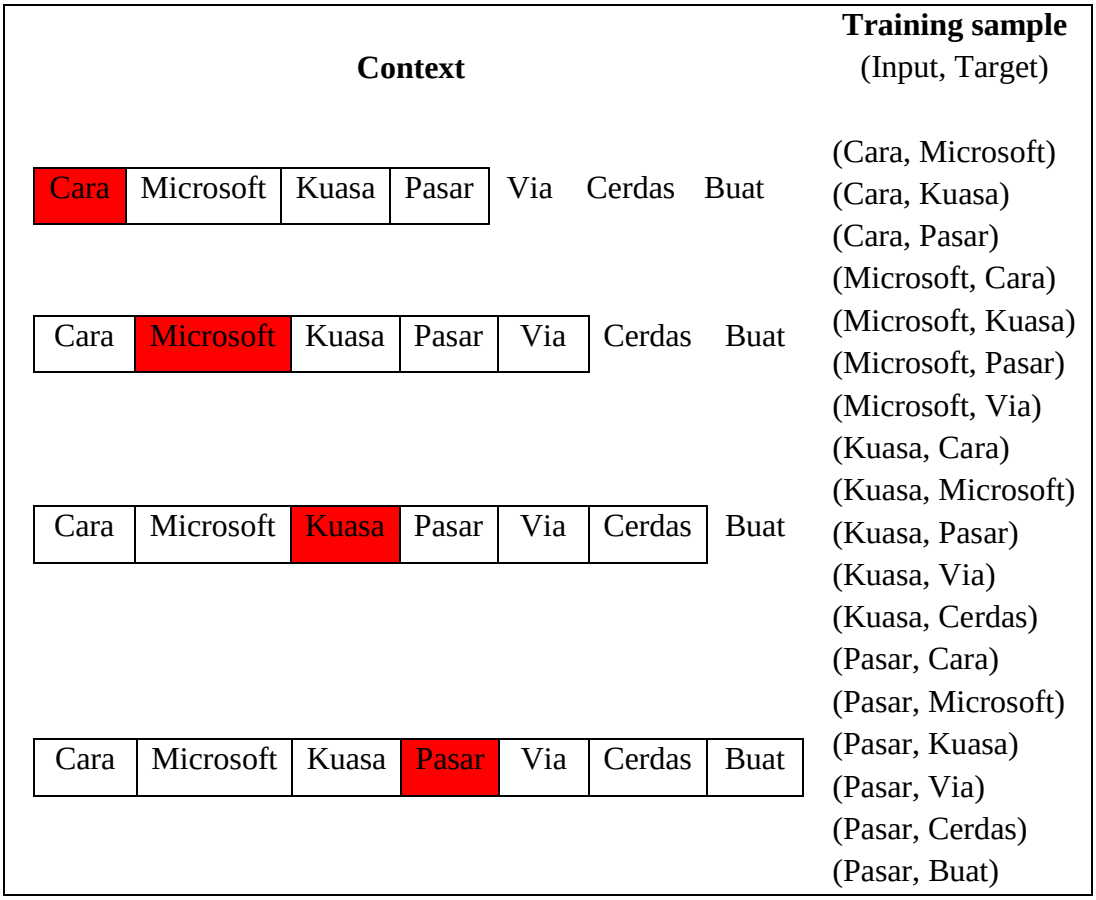
Tabel 2. 7 Data Pembobotan Word2Vec

Kode	Teks
D1	peran guru tetap penting dalam didik ai bukan ganti lain alat dapat bantu memberi alam belajar lebih baik kombinasi antara cerdas manusia buatan adalah kunci berhasil
D2	didik adalah kunci untuk masa depan sukses era digital cerdas buat ai telah main peran penting dalam kaya alam ajar mari kita lihat bagaimana bantu transformasi
D3	bagian guru mulai galau sejak muncul chatgpt dengan cerdas buat ai seperti merasa bantu kaligus ancam eksistensi bagaimana benar peran dalam dukung didik ajar
D4	teknologi ai para pedofil guna cerdas buat untuk materi pelecehan seksual anak para pedofil teknologi jual materi hingga lihat nyata demikian temu
D5	cara microsoft kuasa pasar via cerdas buat

Tabel 2. 8 Indeks Korpus Per-kata Word2Vec

Kode									
D1		D2		D3		D4		D5	
Kosakata	Indeks	Kosakata	Indeks	Kosakata	Indeks	Kosakata	Indeks	Kosakata	Indeks
peran	1	didik	1	bagian	1	teknologi	1	cara	1
guru	2	adalah	2	guru	2	ai	2	microsoft	2
tetap	3	kunci	3	mulai	3	para	3	kuasa	3
penting	4	untuk	4	galau	4	pedofil	4	pasar	4
dalam	5	masa	5	sejak	5	guna	5	via	5
didik	6	depan	6	muncul	6	cerdas	6	cerdas	6
ai	7	sukses	7	chtgpt	7	buat	7	buat	7
bukan	8	era	8	dengan	8	untuk	8		
ganti	9	digital	9	cerdas	9	materi	9		
lain	10	cerdas	10	buat	10	pelecehan	10		
alat	11	buat	11	ai	11	seksual	11		
dapat	12	ai	12	seperti	12	anak	12		
bantu	13	telah	13	merasa	13	para	13		
memberi	14	main	14	bantu	14	pedofil	14		
alam	15	peran	15	kaligus	15	teknologi	15		
belajar	16	panting	16	ancam	16	jual	16		
lebih	17	dalam	17	eksistensi	17	materi	17		
baik	18	kaya	18	bagaimana	18	hingga	18		
kombinasi	19	alam	19	benar	19	lihat	19		
antara	20	ajar	20	peran	20	nyata	20		
cerdas	21	mari	21	dalam	21	demikian	21		
manusia	22	kita	22	dukung	22	temu	22		
buatan	23	lihat	23	dikit	23				
adalah	24	bagaimana	24	ajar	24				
kunci	25	bantu	25						
berhasil	26	tranformasi	26						

Secara Arsitektur Skip-Gram menggunakan current word (sebagai input) untuk memprediksi konteks (sebagai target) disekitarnya, dimana Skip-Gram akan mempelajari distribusi probabilitas dari kata-kata didalam konteks dengan windows yang telah di tentukan. Misal konteks yang digunakan saat ini adalah "Cara Microsoft Kuasa Pasar Via Cerdas Buat" dengan nilai windows = 3.

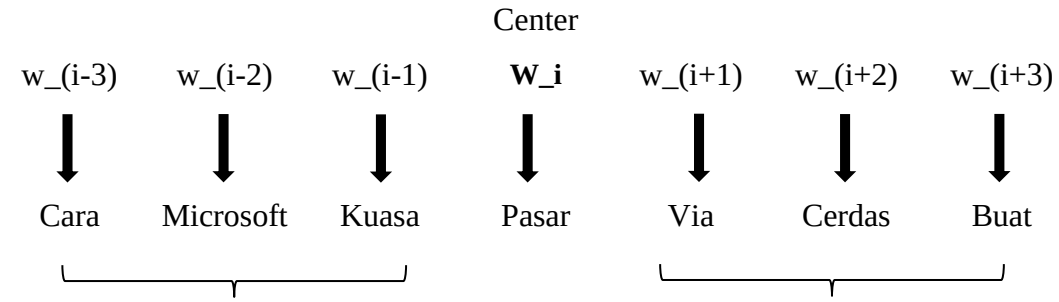


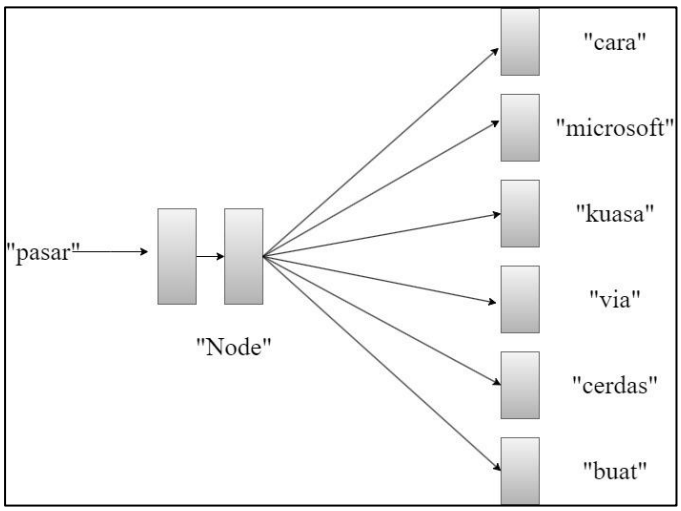
Gambar 2. 4 Contoh Data Training

Untuk merepresentasikan konteks kedalam arsitektur Skip-Gram, maka kita harus merubah setiap kata menjadi one-hot encoded vectors.

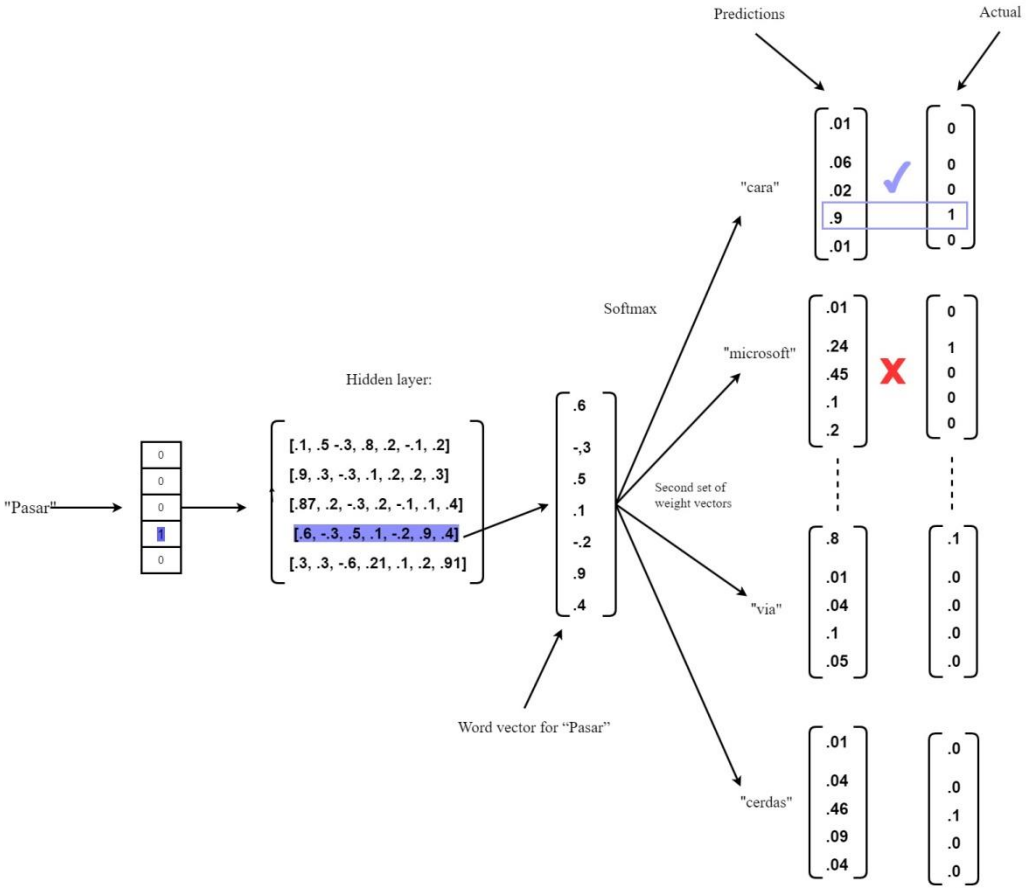
Cara	= [1,0,0,0,0,0,0]	Via	= [0,0,0,0,1,0,0]
Microsoft	= [0,1,0,0,0,0,0]	Cerdas	= [0,0,0,0,0,1,0]
Kuasa	= [0,0,1,0,0,0,0]	Buat	= [0,0,0,0,0,0,1]
Pasar	**= [0,0,0,1,0,0,0]**		

Kemudian, kita akan mengalikan vektor One-Hot Encoding ini dengan matriks bobot Wi, yang memiliki





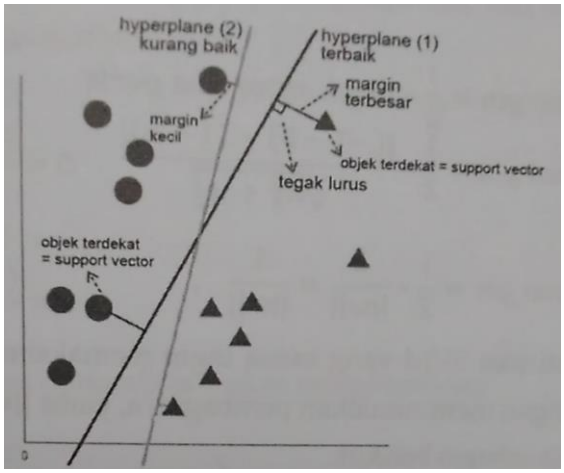
Goal : Hitung probabilitas setiap kata konteks muncul



2.8 *Support Vector Machine*

Support Vector Machine (SVM) merupakan sebagai salah satu algoritma pembelajaran mesin terawasi yang paling efektif. Ini mampu memberikan hasil yang mengesankan tanpa memerlukan penyesuaian ekstensif ketika diterapkan pada masalah tertentu. Namun, terkadang hal ini dianggap sebagai kotak hitam karena landasan matematika yang kuat. SVM telah dikembangkan selama bertahun-tahun melalui upaya kolaboratif beberapa individu. Algoritma SVM awal diperkenalkan oleh Vladimir Vapnik pada tahun 1963. SVM telah menunjukkan keberhasilan penggunaan dalam tiga domain utama: kategorisasi teks, pengenalan gambar, dan bioinformatika(Saraswat, 2023).

Dalam konteks Support Vector Machine, istilah "vektor" memiliki arti yang penting. Vektor adalah entitas matematika yang dapat divisualisasikan sebagai panah, Support Vector Machines (SVM) bertujuan untuk menemukan hyper plane yang optimal. Hyperplane optimal ini adalah hyperplane yang paling baik dalam memisahkan data(Saraswat, 2023).



Gambar 2. 5 Mencari fungsi pemisah optimal untuk obyek yang linearly .separable

Margin adalah jarak antara hyperplane dengan data terdekat dari setiap kelas. Data terdekat ini disebut support vector. Hyperplane adalah yang terbaik pemisah antara dua kelas yang telah ditentukan. Prinsip dasar SVM adalah pengklasifikasi linier, kemudian dikembangkan agar dapat bekerja pada masalah non-linier, yaitu dengan memasukkan konsep trik kernel pada ruang kerja berdimensi tinggi. Kernel SVM yang digunakan dalam penelitian ini adalah kernel Linear, *Radial Basis Function* (RBF).

Metode SVM memiliki konsep utama dalam mengklasifikasikan data yaitu menemukan hyperplane terbaik untuk memisahkan antara dua kelas yang telah ditentukan. Hyperplane terbaik diperoleh dengan memaksimalkan margin support vector. Proses memaksimalkan margin support vector dapat dilakukan dengan

meminimalkan Lagrangian dan menurunkannya dari w dan b ditemukan pada persamaan 1 dengan kondisi 1 dan kondisi 2(Styawati. *et al.*, 2021).

Lp = ||w||2 – i yi (w.xi + b) – 1 (1)

Kondisi 1:

W = i yi $\sum_{i=1}^{N} a$ (2)

Kondisi 2:

B = yi – w.xi (3)

Dalam proses memaksimalkan pengali Lagrangian, masih banyak kemungkinan nilai w, b, dan. Berdasarkan permasalahan tersebut, maka proses memaksimalkan pengali Lagrange harus ditransformasikan ke dualitas pengali Lagrange pada persamaan 5 dengan kendala 1 dan 2.

Max $Ld = \sum_{i=1}^{N} a_i - \frac{1}{2} \sum_{i,j}^{N} a_i \ a_j \ y_i \ y_j \ x_i \ x_j$ (4)

Kondisi 1:

$\sum_{i=1}^{N} a_i$ yi = 0, I = 1,2,…..,N (5)

Komdisi 2:

0<= αi<=C, i=1, i = 1,2 …,N (6)

Namun SVM juga memiliki banyak kedala. Kendala pertama adalah everfitting. Perlu mencari fungsi kernel yang tepat dan aspek-aspek lain untuk meningkatkan regularization. Selain itu SVM tidak menyediakan nilai probabilitas dari klasifikasi yang dihasilkan. SVM pada awalnya diperuntukan untuk mrngklasifikasi dua kelas . Namun perkembangan selanjutnya, Multi-Class SVM banyakdigunakan. Salah satu yang sering digunakan adalah prinsip “one-against-one” yang diperkenalkan oleh Knerr tahun 1990(Handayanto and Herlawati, 2020).

2.9 Klasifikasi SVM

Klasifikasi adalah proses pengelompokan banyak data ke dalam kelas-kelas yang telah ditentukan dan diberikan menurut kesamaan ciri dan pola yang terkandung dalam kata-kata tersebut. Secara umum, proses klasifikasi diawali dengan penyediaan data apa saja yang dijadikan acuan untuk membuat aturan klasifikasi data. Data tersebut biasanya dikenal sebagai set pelatihan. Dari set pelatihan, kemudian dibuat model untuk mengklasifikasikan data. Model tersebut kemudian digunakan sebagai acuan untuk mengklasifikasikan kelas data yang tidak diketahui yang dikenal dengan *test set*. Metode klasifikasi yang akan digunakan pada penelitian ini adalah SVM.

Support Vector Machine (SVM) merupakan salah satu metode Machine *Learning* untuk *Pattern Recognition* yang sering dugunakan dalam penelitian *Text Recognition. Support Vector Machine* dapat diterapkan dalam teknik melakukan sebuah prediksi, pada kasus klasifikasi ataupun regresi. Konsep dasar dari *Support Vector Machine* adalah mencari *Hyperplane* yang memaksimumkan margin. *Hyperplane* dapat berupa *line* pada *two dimension* atau dua kelas dan dapat berupa *flat plane* pada *multiple dimension* atau multikelas. Margin adalah jarak antara hyperplane dan data terdekat dari masing-masing kelas. Data terdekat disebut support vector. Hyperplane adalah pemisah terbaik antara dua kelas yang telah ditentukan. Prinsip dasar SVM adalah pengklasifikasi linier dan kemudian dikembangkan untuk mengerjakan masalah non linier. Dengan memasukkan konsep trik kernel dalam ruang kerja berdimensi tinggi. Kernel SVM yang digunakan dalam penelitian ini adalah kernel RBF untuk proses transformasi dari input space menjadi feature space(Styawati and Mustofa, 2019).

Jaringan radial basis function (RBF network) memiliki model jaringan yang hampir menyerupai metode jaringan syaraf tiruan *multilayer perceptron* (MLP network). Jaringan RBF suatu jaringan yang memiliki dua layer. Ada dua perbedaan antara RBF dan dua layer pada jaringan perceptron. Pada layer pertama dari jaringan RBF tidak menggunakan operasi perkalian antara bobot dan input (perkalian matriks), tetapi menggunakan perhitungan jarak antara vektor input dan baris dari bobot matriks yang mana hal ini mirip dengan metode jaringan syaraf tiruan *Learning Vector Quantization* (LVQ network), dan kedua tidak menambahkan nilai bias. Adapun tahap pelatihan dengan menggunakan algoritma RBF adalah sebagai berikut:

a. Inisialisasi bobot (set nilai secara acak).

b. Lakukan c – h sampai berhenti.

c. Untuk masing-masing input lakukan langkah d – g.

d. Masing-masing input x*i*, *i* = 1, 2, 3, … n dihitung secara keseluruhan.

e. Hitung fungsi aktivasi Gaussian jaringan RBF dengan menggunakan persamaan berikut.

$$\varphi(r) = exp\left(-\frac{r^2}{2\sigma^2}\right)$$

Dimana (r) = fungsi aktivasi Gaussian, r =input, dan σ = nilai spread.

f. Hitung output keseluruhan jaringan RBF dengan mengunakan persamaan berkut.

$$Y_{net} = \sum_{i=1}^{H} w_{im}\varphi_{i(r)} + w_0$$

Dimana Y_{net} = output jaringan, $\varphi_{i(r)}$ = nilai fungsi aktivasi.

Pada Penelitian ini yang berjudul Analisis pengujan pembelajaran terhadap jaringan radial basis function (RBF network) diperoleh pembalajaran yang baik yaitu 95%, karena perhitungan iterasi yang cepat dengan menggunakan perhitungan matriks Gaussian dan jaringan yang hampir menyerupai dengan model jaringan *multilayer prceptron*(Azmi, 2016).

Metode SVM memiliki konsep sentral dalam mengklasifikasikan data yaitu mencari hyperplane terbaik untuk memisahkan antara dua kelas yang telah ditentukan. Hyperplane terbaik diperoleh dengan memaksimalkan margin support vector. Proses memaksimalkan support vector margin dapat dilakukan dengan meminimalkan lagrangian dan direduksi menjadi w dan b yang terdapat pada persamaan 1 dengan suku 1 dan 2.

$$Lp\,\frac{1}{2}\|w\|^2 - \sum_{i=1}^{N} a_i\, y_i = (\, w.x_i + b\,) - b$$

Ketentuan 1:

$$w = \sum_{i=1}^{N} a_i\, y_i\; x_i$$

Ketentuan 2:

$$b = y_i - w.x_1$$

Karena nilai α tidak diketahui, nilai w dan b tidak dapat ditentukan. Nilai α dicari dengan memaksimalkan pengali Lagrangian dengan kondisi optimal untuk dualitasnya menggunakan kendala *Karush-Kuhn-Tucker* (KKT). Penggunaan batasan KKT menjadikan nilai pengali *Lagrange* (α) sama dengan jumlah data latih. Proses pemaksimalan pengali Lagrangian masih memiliki banyak kemungkinan nilai w, b, dan α. Berdasarkan permasalahan tersebut, proses

maksimisasi pengali Lagrange harus ditransformasikan menjadi dualitas pengali Lagrange pada persamaan 5 dengan batasan 1 dan 2.

$$\text{Maks } Lp = \sum_{i=1}^{N} a_i \; - \frac{1}{2} = \sum_{i.j}^{N} \; a_i \, a_j \, y_i \, y_j \, x_{i.} x_j$$

Ketentuan 1:

$$\sum_{i=1}^{N} \; a_i \, a_j = 0_{,i} \, i = 1,2, \dots, N$$

Ketentuan 2:

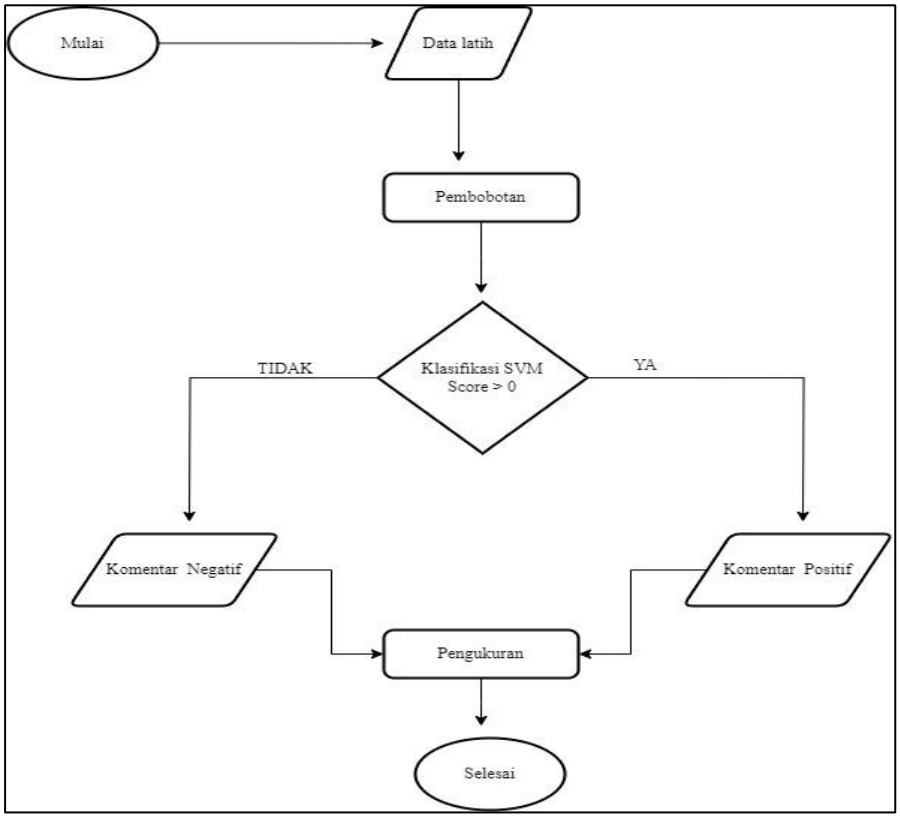
$$0 \leq ai \leq C, i = 1,2, \dots, N$$

Setelah didapatkan nilai w, b, dan α, selanjutnya ditentukan label dengan menggunakan model

$$F\big(\Phi(x)\big) = sign \; (w.\Phi(x) + b)$$

Jika nilai dari *f(x)* dihasilkan adalah*f(x)>0*maka data tersebut diklasifikasikan ke dalam kelas positif (+1), jika *f(x)<0* maka data tersebut diklasifikasikan ke dalam kelas negatif (-1).

Berikut ini gambar 3.9 adalah proses penerapan dari metode *Support Vector Machine two dimension.*



Gambar 2. 6 Proses Klaifikasi SVM dua dimensi

Berikut ini adalah contoh perhitungan manual SVM :

Tabel 2. 9 Contoh Perhitungan Manual SVM

X_1	X_2	Y (Kelas)
1	1	1
1	-1	-1
-1	1	-1
-1	-1	-1

Terdapat dua fitu yaitu X_1 dan X_1 sehingga otomatis w (bobot) aka nada dua (W_1 dan X_1). Langkah selanjutnya adalah meminimalkan nilai *margin* dengan rumus sebagai berikut :

$$\frac{1}{2}\|W\|^2 = (\frac{1}{2}W_1^2\ W_2^2\ )$$

Dengan Syarat : $y_1(x_1.w + b)$ -1 ≥ 0

$y_1(x_1.w1 + x_2\ w2 + b) \geq 1$

I = 1,2,3……,n

Setelah itu masukan data kedalam syarat tersebut, sehingga menjadi :

Persamaan 1 : $(w_1 + w_2 + b) \geq 1$ $\rightarrow y_1 = 1.x_1 = 1.x_2 = 1$

Persamaan 2 : $(-w_1 + w_2 - b) \geq 1$ $\rightarrow y_2 = -1.x_1 = 1.x_2 = -1$

Persamaan 3 : $(w_1 - w_2 - b) \geq 1$ $\rightarrow y_3 = -1.x_1 = -1.x_2 = 1$

Persamaan 4 : $(w_1 + w_2 - b) \geq 1$ $\rightarrow y_4 = -1.x_1 = -1.x_2 = -1$

Setelah ke empat persamaan di eliminasi, maka akan menjumlahkan persamaan

Persamaan (1) dan (2) = $(w_1 + w_2 + b) \geq 1 + (-w_1 + w_2 - b) \geq 1 = 2\ w_2 = 2$

$w_2 = 1$

Persamaan (1) dan (3) = $(w_1 + w_2 + b) \geq 1 + (w_1 - w_2 - b) \geq 1 = 2\ w_1 = 2$

$w_1 = 1$

Persamaan (2) dan (3) = $(-w_1 + w_2 - b) \geq 1 + (w_1 - w_2 - b) \geq 1$ = -2b = 2

b = -1

Selanjutnya memvisualisasikan garis *Hyperlane* dengan rumus :

Hasil persamaan hyperplane

$(w_1.x_1 + w_2.x_2 + b = 0) \rightarrow 1.x_1 + 1.x_2 + -1 = 0 \rightarrow x_1 + x_2 - 1 = 0 \rightarrow x_1 = 1 - x_1$

Tabel 2. 10 Hasil persamaan hyperplane

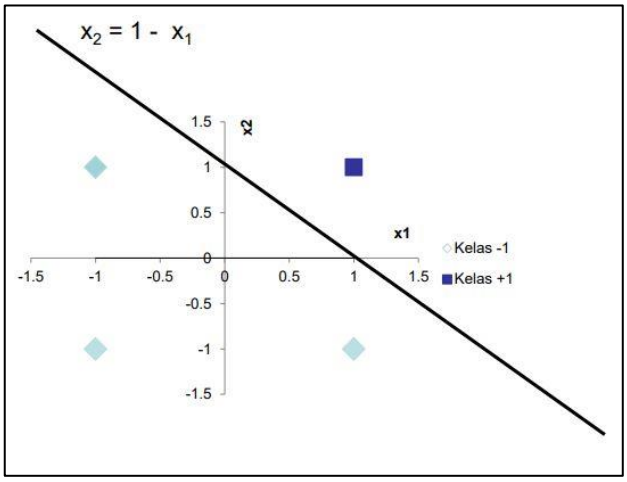
X_1	-2	-1	0	1	2
$X_2 = 1 - X_1$	3	2	1	0	-1

Diketahui data uji : $f(x) = X_1 + X_2 - 1$
Kelas = sign(f(x))

Tabel 2. 11 Data Uji Perhitungan manual SVM

Data Uji		Hasil Klasifikasi
X_1	X_2	Kelas = sign ($X_1 + X_2 - 1$)
0	5	Sign (0+5-1) = +1
-1	3	Sign (-1+3-1) = +1
6	-1	Sign (6+(-1)-1= +1
2	-3	Sign (2+(-3)-1)= -1
3	-7	Sign (3+(-7)-1) = -1

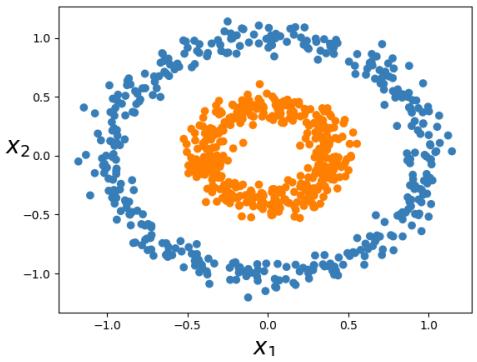
Sehingga garis *Hyperplane* akan melalui titik (x_1, x_1) yaitu (1,1). Jika diivisualisasikan maka *Hyperplane* akan menjadi seperti dibawah ini.



Gambar 2. 7 Visualisasi garis Hyperplane

2.10 Metode Kernel

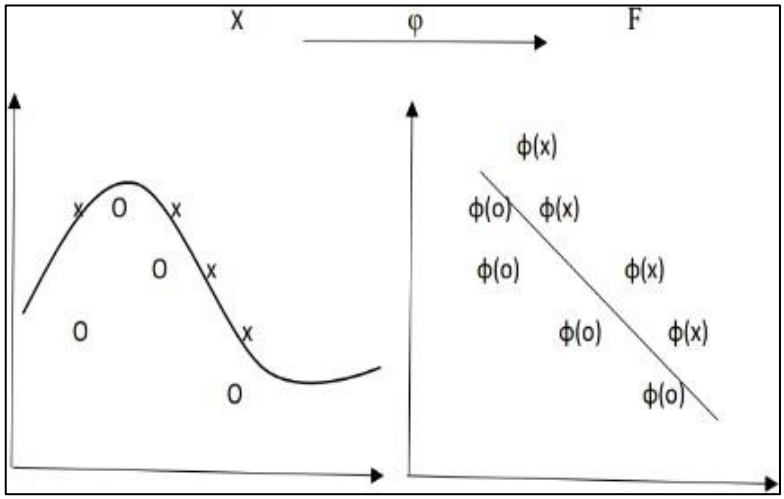
Banyak teknik data mining atau machine learning yang dikembangkan dengan asusmsi keliniearan. Sehingga algoritma yang dihasilkan terbatas untuk kasus-kasus yang linier. Karena itu, bila suatu kasus klasifikasi memperlihatkan ketidak linearan, algoritma yang di desain berdasarkan asumsi kelinearan, seperti perceptron misalnya, tidak bisa mengatasinya. Secara umum kasus-kasus yang terjadi di dunia nyata bukanlah kasus yang tidak linier. Sebagai contoh, perhatikan Gambar 2.1. Data ini sulit dipisahkan secara linear. Metode kernel(Santosa and Umam, 2018) adalah salah satu untuk mengatasinya. Dengan metode kernel, suatu data *x* di input space di-*mapping* ke *feature space F* dengan dimensi yang lebih tinggi melalui map φ sebagai berikut : x → φ (*x*). Karena itu data x di input space menjadi φ (x) di *feature space*.



Gambar 2. 8 Ilustrasi data dua kelas yang tidak linear

Sering kali fungsi $\emptyset(x)$ tidak tersedia atau tidak bisa dihitung. Tetapi *dot product* dari hasil dua vector dapat dihiktung baik didalam *input space* maupun di *feature space.* Dengan kata lain, sementara $\emptyset(x)$ mungkin tidak diketahui, *dot product* $<\emptyset(x1), \emptyset(x2)>$ masih bisa dihitung di *feature space.* Untuk bisa memakai metode kernel, pembatas (*constraint)* perlu diekspresikan dalam bentuk

dot product dari vector data. Sebagai konsekuensi, pembatas yang menjelaskan permasalahan dalam klasifikasi harus diformulasikan kembali sehingga menjadi bentuk *dot product*. Dalam *feature space* ini, dot product <.> menjadi <∅*(x),* ∅*(x)*>. Suatu fungsi kernel, *k(x, x'),* bisa untuk menggantikan *dot product* <∅*(x1)* ∅*(x2)*>. Kemudian di *feature space,* kita bisa membuat suatu fungsi pemisah yang linear yang mewakili fungsi *non-linear* di *input space.* Gambar 2.2 mendeskripsikan suatu contoh *feature mapping* dari ruangan dua dimensi ke feature space dua dimensi. Dalam *input space*, data tidak bisa dipisahkan secara linear, tetapi kita bisa memisahkan di *feature space.* Karena itu, dengan memetakan data ke *feature space* ini, menjadi tugas klasifikasi menjadi lebih mudah(Santosa and Umam, 2018).



Gambar 2. 9 Kernel Map Mengubah Persoalan Yang Tidak Linear Menjadi Linear Dalam Space Yang Baru (Suyanto 2018)

Berikut adalah beberapa fungsi kernel yang biasanya dipakai dalam literatur SVM (Haykin, 1999):

Kernel linear (x,y) : $x^T.y$

- Kernel linear (x,y) adalah nilai kernel linear antara dua vektor fitur x dan y
- x^T adalah transpos dari vektor x

Kernel polynominal : $(X^T X_i + 1)^p$	• . adalah operator dot product (produk dot) • x dan x_i adalah vektor fitur dari dua sampel data • p adalah orde (derajat) polynomial yang ditentukan sejauh mana fitur-fitur dipertinggi dalam ruang fitur tinggi
Kernel Radial baris function (RBF) : $\exp(- \frac{1}{2\sigma^2}\|x-x_i\|^2)$	• $K_{RBF}(x,x_i)$ adalah nilai kernel RBF antara dua vector fitur x dan x_i • Exp adalah fungsi eksponensial • σ (sigma) adalah parameter yang disebut sebagai lebar kernel, yang mengontrol berapa jauh pengaruh setiap titik data terhadap yang lain
Kernel tagent hyperbolic : tanh $(\beta^T X_i + \beta_i)$	• $K_{tanh}(x,xi)$ adalah nilai kernel tangent hyperbolic antara dua vektor fitur x dan x_i • tanh adalah fungsi tangen hiperbolik • β adala vector bobot atau parameter • X_i adalah vector fitur dari sampel data i • β_0 adalah bias atau konstanta

2.11 Confusion Matrix

Confusion Matrix adalah teknik yang digunakan untuk mengevaluasi klasifikasi model untuk memperkirakan objek yang benar atau salah. Sebuah matriks dari prediksi akan dibandingkan dengan kelas asli yang berisi informasi aktual dan prediksi nilai klasifikasi. Setelah sistem berhasil melakukan klasifikasi tweet, dibutuhkan ukuran untuk menentukan seberapa valid atau tepat klasifikasi yang telah dibuat oleh sistem. Tabel 2.1 ini akan menunjukan confusion matrix yang digunakan untuk membantu dalam perhitungan system evaluasi(Pravina, Cholissodin and Adikara, 2019).

Pengujian akurasi ini dilakukan menggunakan confusion matrix dengan melibatkan empat kondisi sebagai berikut:

Tabel 2. 12 Confusion Matrix

Classification	*Predicted Positive*	*Predicted Negatif*
Actual Negative	*True Negative (TN)*	*False Positive (FP)*
Actual Positive	*False Negative (FN)*	*True Positive (TP)*

Dalam pengukuran kinerja *Confusion Matrix* terdapat empat bagian untuk mengidentifikasi suatu prediksi, berikut diantaranya(Santosa and Umam, 2018):

1. TN (*True Negative*) adalah output kelas negatif yang berhasil ditebak sebagai kelas negatif.
2. TP (*True Positive*) adalah output kelas positif yang berhasil ditebak sebagai kelas positif.
3. FN (*False Negative*) adalah output kelas positif yang berhasil ditebak sebagai kelas negatif.
4. FP (*False Positive*) adalah output kelas negatif yang berhasil ditebak sebagai kelas positif.

Hasil dari pengujian *Confusion Matrix* dapat menghasilkan nilai Accuracy, Recall, Precision. Pada klasifikasi biner terdapat beberapa nilai evaluasi yang sering digunakan. Dapat dilihat berdasarkan nilai *confusion matrix*

1. *Accuracy* (ACC) adalah efektivitas keseluruhan dari hasil klasifikasi.

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

2. *Precision* (PREC) merupakan presentase dari label data dengan label positif yang diberikan oleh klasifikasi.

   Precision (%) = $\frac{FN}{(FP+FN)}$

3. *Recall* (REC) atau sensitivity adalah efektivitas dari pengklasifikasi dalam mengidentifikasi label positif.

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