BAB II
LANDASAN TEORI

2.1 Penelitian Sebelumnya

Penelitian sebelumnya bertujuan sebagai bahan acuan dalam menyusun penelitian ini, sehingga dapat memberikan gambaran mengenai proses dan hasil yang relevan dengan penelitian yang akan dilakukan penulis.

Penelitian ini membahas penerapan metode *SVM* dalam menganalisis sentimen *review* pelanggan hotel pada Tripadvisor.com. Metode ini digunakan karena keefektifanya dibadingkan klasifikasi sejenis seperti *Artifical Neural Network (ANN)* khususnya dalam menemukan solusi dan metode ini baik dalam penyelesaian analisis. Adapun penggunaan *Term Frequency-Inverse Document Frequnecy (TD-IDF)* sebagai pembobotan kata. Perangkat lunak untuk memproses data menggunakan *Python*. Dalam melakukan penelitian ini menggunakan 223 data ulasan, menghasilkan kelas *positif* sebanyak 177 ulasan, kelas *negative* sebanyak 46. Tingkat akurasi yang diperoleh sebesar 88% [9].

Penelitian selanjutnya yaitu mengoptiasi *SVM (Support Vector Machine)* berbasis *PSO (Particle Swarm Optimazation)* untuk menganalisis sentimen perpindahan ibu kota Indonesia di *Twitter*. Perangkat lunak yang digunakan yaitu *Python*. Pemilihan metode *SVM* khususnya untuk menganalisis sentimen merupakan metode terbaik dan memiliki keunggulan dalam komputasi data berdimensi tinggi, selain itu penggunaan *feature selection PSO* untuk meningkatkan akurasi *SVM* dalam menganalisis sentimen perpidahan ibu kota Indonesia. kata kunci untuk *mengcrawling* data yaitu *#PindahIbuKota*, *#IbuKotaPindah* dan *#IbuKotaBaru*. Penelitian ini menggunakan data sebesar 1.319 *tweets* dengan *sentimen positif* 457, *sentimen negative* 862. Nilai akurasi yang dihasilkan *SVM* semula sebesar 79.06% dan ditingkatkan oleh *PSO* sebesar 2.09% sehingga total akurasi sebesar 81.15% [10].

Penelitian selanjutnya dilakukan oleh Ratino dkk, membandingkan kinerja algoritma *Naive Bayes Classification* dan *Support Vector Machine (SVM)* dalam meganalisis sentimen informasi *Covid-19*. Media sosial instagram menjadi

persebaran informasi mengenai pandemi *Covid-19* yang sedang melanda di dunia. Analisis sentiment merupakan kumpulan *teks opini* atau ulasan dari pengguna dalam menanggapi teks, video, gambar. Tujuan penelitian ini untuk mengetahui repson pengguna di social media yaitu, Instagram dalam menanggapi *Covid-19* dengan menggunakan bahasa inggris dan membandingkan kinerja *Algoritma Naive Bayes* dan *SVM*. Perangkat lunak yang digunakan yaitu *Rapid Miner*. Penelitian ini menghasilkan akurasi *Naive Bayes* sebesar 78,02%, ditingkatkan menggunakan *Particle Swarm Optimazation* menjadi 79,07%. Hasil akurasi menggunakan *Support Vector Machine* sebesar 80,23% ditingkatkan menggunakan *Particle Swarm Optimazation (PSO)* menjadi 81,16% , dapat disimpulkan bahawa *SVM-PSO* lebih unggul dari pada *Naïve Bayes –PSO* [18].

Penelitian selanjutnya dilakukan oleh Wirdhayanti Paulina dkk, menganalisis sentimen ulasan pelanggan terhadap *Guest House* menggunakan metode *Support Vector Machine*. Premium Guest House merupakan salah satu tempat penginapan di Kota Malang yang menyediakan *E-WOM (Electronic Word of Mouth)* untuk mengetahui respon pelanggannya terhadap pelayanan yang telah diberikan. Penyedia ulasan yang diberikan dibagi menjadi dua yaitu *Guest Review* (manual menggunakan kertas) dan ulasan online merupakan data yang akan dijadikan bahan penelitian melalui situs *Online Travel Agent (OTA)*. Hasil dari proses *scrapping* menghasilkan data sebesar 1561 data teks ulasan data dari lima situs yaitu, Agoda.com, Expedia.com, Pegi-pegi.com, Booking.com, TripAdvisor.com dari tahun 2012 sampai 2019. Analisis sentimen yang digunakan terdiri dari 3 aspek yaitu, kamar diperoleh 1070 ulasan (860 data positif dan 210 data negatif), harga diperoleh 461 ulasan (414 data positif dan 47 data negative) dan layanan diperoleh 885 ulasan (816 data positif dan 69 data negative). Hasil penelitian menunjukan aspek layanan memiliki akurasi tertinggi sebesar 0.83, akurasi aspek kamar sebesar 0.79 dan akurasi terendah yaitu harga sebesar 0.70 [19].

Penelitian selanjutnya dilakukan oleh Iman Nur Fakhri dkk, dalam menganalisis sentimen kepuasan layanan dan fasilitas di Universitas Telkom menggunakan *Support Vector Machine (SVM)* dan pembobotan *Term Frequency - Invers Document Frequency (TF-IDF)*. Analisis sentimen ini bertujuan untuk

meningkatkan mutu dan kualitas layanan yang diberikan kepada mahasiswa. Data yang digunakan dibagi menjadi 3, berdasarkan banyaknya data yaitu, 10.000, 7000 dan 3000 data dengan penggunaan 3 kernel yaitu, *kernel linear*, *polynomial*, dan *RBF*. Jumlah data testing yang digunakan sebanyak 10.000 data. Berdasarkan penelitian yang telah dilakukan akurasi tertinggi di capai sebesar 69,7% dengan data sebanyak 10.000 menggunakan kernel linear, dengan 67% sentimen positif dan 33% sentimen negative. Nilai akurasi sebesar 69.7%% menggunakan *Karnel Linier* [20].

Penelitian selanjutnya dilakukan oleh Valentino Kevin Sitanayah Que dkk, dalam menganalisis sentimen transportasi online menggunakan *Support Vector Machine (SVM)* dan *Particle Swarm Optimazation (PSO)*. Perkembangan transportasi *online* di Indonesia cukup meningkat karena kemudahan dalam memperoleh transportasi tanpa harus menunggu lama. Analisis sentimen ini bertujuan untuk mengetahui respon pengguna media sosial yaitu twitter dalam memberikan opini mereka atas pelayanan dan dapat mengetahui perbandingan antara akurasi *SVM* dan *SVM-PSO* dengan *tool* yang digunakan yaitu *RapidMiner*. Hasil penelitian menggunakan *SVM* menunjukan ulasan bersifat *positif* sebesar 702 data atau 62% dan ulasan *negatif* sebesar 48% atau 428 data. Selanjutnya untuk analisis menggunakan *SVM-PSO* yaitu ulasan *positif* sebesar 53% atau 604 data dan ulasan *negatif* sebesar 526 data atau 47%. Perbandingan jumlah baik ulasan positif dan ulasan negatif sangat mempengaruhi tingkat akurasi yang diperoleh yaitu *SVM* dengan akurasi 95,46% dan *SVM-PSO* menghasilkan akurasi sebesar 96,04% [21].

Penelitian selanjutnya dilakukan oleh Rian Tineges dkk, dalam menganalisis sentimen untuk pelayanan indihome di twitter menggunakan metode klasifikasi yaitu *Support Vector Machine (SVM)* dan *tool* yang digunakan yaitu *python*. Pengguna intenet yang semakin meningkat, hampir 64,8% masyarakat Indonesia merupakan pengguna internet. Berbagai informasi seperti bisnis, hiburan, ekonomi dan politik ada di *twitter*. Indihome termasuk salah satu operator penyedia ISP yang paling banyak digunakan masyarakat Indonesia sebesar 8,7%. Tujuan penelitian ini yaitu untuk mengetahui respon pengguna layanan indihome dengan melakukan analisis sentimen di twitter. Adapun *keyword* yang digunakan dengan menyebutkan

username @IndiHome. Data yang digunakan sebanyak 1400 data dalam jangka waktu 16 sampai 23 MAret 2020, olah data dilakukan dengan 80% atau 1120 tweet untuk *data training* dan data uji 20% atau 280 tweet. Berdasarkan penelitian yang dilakukan menghasilkan 104 data atau 18,4% merupakan sentimen positif dan 176 data atau 81,6% merupakan sentimen negatif dan akurasi yang dihasilkan sebesar 87% [22].

Penelitian ini dilakukan oleh Putu Mega Nirmala Dharmapatni dkk, dalam menerapkan *Algoritma Support Vector Machine* untuk menganalisis kenaikan tarif BPJS di *Twitter*. Badan Jaminan Kesehatan dan Sosial (BPJS) merupakan Badan Usaha Milik Negara (BUMN) yang menyediakan layanan layanan kesehatan untuk seluruh masyarakat Indonesia. Tujuan dari penelitian ini yaitu untuk mengetahui *feedback* atau umpan balik sebagai respon masyarakat dalam hal kenaikan tarif yang telah di berikan oleh BPJS melalui *Twitter*. *SVM* digunakan sebagai metode klassfikasi sentimen analisis yang baik untuk data berjumlah besar. Data yang diambil mulai dari bulan Januari 2020 sampai bulan Juni 2020, dengan menggunakan *Application Programming Interface (API).* Data yang didapatkan selanjutnya akan dibagi mejadi dua bagian yaitu data *testing 20%* dan data *training 80%*. Adapun *keyword* yang digunakan yaitu, *#KenaikanBPJS*. Berdasarkan 671 data, terdapat 271 yang menyetujui kenaikan tarif BPJS dan 400 data yang kurang setuju adanya kenaikan tarif BPJS. Adapun akurasi yang dihasilkan sebesar 92% (87*% sentimen positif* dan 96% *sentimen negatif*)[23].

Adapun tinjauan pustaka dalam bentuk tabel yang dapat dilihat pada tabel 2.1

Tabel 2.1 Kajian Pustaka

No	Judul	Objek	Metode	Masalah	Hasil	Perbedaan
1	Penerapan Metode *Support Vector Machine* untuk Analisis Sentimen pada Review Pelanggan Hotel[9]	Sentimen hotel di Tripadvisor.com	SVM ini di perlukan karena metode ini digunakan untuk menyelesaikan masalah klasifikasi genexpression analysis.	Perkembangan internet yang semakin pesat membuat banyak orang mengakses internet untuk mendapatkan berbagai macam informasi, salah satunya dalam mencari informasi ulasan hotel. Analisis sentimen ini diperlukan untuk mengetahui	Hasil penelitian menunjukkan bahwa *SVM* dapat membantu dalam menyelesaikan analisis sentiment hotel di situs Tripadvisor.com dengan akurasi yang dihasilkan sebesar 88%, dan data ulasan yang diteliti sebanyak 223 data, 176 positif dan 46 negatif.	Perbedaan penelitian sebelumnya menggunakan objek penelitian pada situs tripadvisor.com dengan metode SVM, sedangkan penelitian ini menggunakan objek penelitian pada situs website agoda.com dan tiket.com menggunakan metode SVM dan PSO untuk

No	Judul	Objek	Metode	Masalah	Hasil	Perbedaan
				respon pelanggan hotel. Semakin banyaknya hotel yang dapat dituju saat bepergian ke luar kota dapat membuat calon pengunjung kebingungan untuk menentukan hotel mana yang cocok untuk mereka kunjungi.		mendapatkan akurasi yang lebih baik, objek penelitian pada situs agoda dan Tiket.com
2	Optimasi SVM Berbasis PSO pada Analisis Sentimen Wacana Pindah Ibu Kota Indonesia[10]	Sentimen Tentang wacana Pindah Ibu Kota Indonesia (*Twitter*)	*SVM* digunakan karena kelebihanya dalam komputasi data	Analisis sentimen ini diperlukan untuk mengetahui opini wacana tentang	Penggunaan *SVM – PSO* dapat membantu peneliti dalam menyelesaikan	Perbedaan penelitian sebelumnya menggunakan objek penelitian pada twitter, sedangkan

No	Judul	Objek	Metode	Masalah	Hasil	Perbedaan
			berdimensi tinggi. PSO digunakan untuk meningkatkan akurasi SVM.	pemindahan ibu kota Indonesia mendapatkan respon dari masyarakat sehingga menjadi *trending topic* pada media sosial *twitter*, namun tidak selalu mendapat respon positif dari pengguna sosial sehingga memicu perdebatan di sosial media twitter.	permasalahan dengan akurasi yang dihasilkan sebelumnya 79,06% kemudian di optimalisasikan menggunakan PSO menjadi 81,15% *(Good Classification)*	penelitian ini menggunakan objek penelitian pada situs website agoda.com dan tiket.com

No	Judul	Objek	Metode	Masalah	Hasil	Perbedaan
3	Sentimen Analisis Informasi Covid-19 menggunakan Support Vector Machine dan Naïve Bayes[18]	Sentimen Tentang Informasi Covid-19 (*Instagram*)	Perbandingan menggunakan algoritma *SVM*(SVM diperlukan karena kelebihannya dalam mengidentifikasi hyperplane terpisah dengan memaksimalkan margin antara kelas terpisah)-*PSO* dan *Naïve Bayes-PSO*. PSO digunakan untuk	Organisasi kesehatan dunia menyatakan COVID-19 terus bermunculan dan merupakan masalah bagi kesehatan masyarakat dunia. Informasi saat ini banyak disampaikan melalui berbagai media sosial salah satunya instagram. Analisis sentimen diperlukan untuk mengetahui opini	Hasil penelitian menunjukkan bahwa baik *Naive Bayes-PSO* dan *SVM-PSO* dapat menyelesaikan permasalahan. Akurasi yang dihasilkan menunjukan *SVM-PSO* lebih unggul dengan akurasi 81,16%.	Perbedaan penelitian sebelumnya menggunakan dua metode untuk menentukan akurasi yang lebih unggul dari kedua metode, sedangkan untuk penelitian ini menggunakan *SVM-PSO*

No	Judul	Objek	Metode	Masalah	Hasil	Perbedaan
			meningkatkan akurasi sebuah algoritma.	masyarakat yang disampaikan melalui komentar pada media sosial instagram terhadap informasi COVID-19.		
4	Analisis Sentimen Berbasis Aspek Ulasan Pelanggan Terhadap Kertanegara Premium Guest House Menggunakan Support Vector Machine[19]	*Guest Review* dan ulasan *online* di situs OTA(*Online Travel Agent*).	*SVM* digunakan karena memiliki efektifitas terbaik diantara algoritme machine learning	Analisis sentimen ini diperlukan untuk mengetahui opini pelanggan yang melakukan perjalanan wisata online selain berfungsi untuk reservasi akomodasi wisata tetapi mempunyai	Hasil penelitian ini menunjukkan bahwa penggunaan *SVM* dan *TF-IDF* dapat menyelesaikan permasalahan dalam menemukan analisis sentiment *Guest House* serta hasil pengujian hasil klasifikasi sentimen	Perbedaan penelitian sebelumnya menggunakan objek penelitian *Guest Review* dan ulasan *online di situs* OTA(*Online Travel Agent)*, sedangkan untuk penelitian ini menggunakan objek penelitian pada

No	Judul	Objek	Metode	Masalah	Hasil	Perbedaan
				peran sebagai media Electronik Word of Mouth(E-WOM) melalui review/ulasan pelanggan dan Kertanegara Premium Guest House sangat menyadari pentingnya keberadaan E-WOM demi kelangsungan bisnis.	memiliki rata-rata yang baik dengan nilai *Accuracy*, *Precision. Recall*, dan F1-*Score* diatas 70%.	agoda.com dan tiket.com
5	Analisis Sentimen pada Kuisioner Kepuasan	Layanan dan Fasilitas di	*SVM* digunakan karena memiliki prisip Structural	Analisis sentimen diperlukan untuk mengetahi opini	Hasil penelitian ini menunjukkan bahwa penggunaan *SVM*	Perbedaan penelitian sebelumnya menggunakan objek

No	Judul	Objek	Metode	Masalah	Hasil	Perbedaan
	Terhadap Layanan dan Fasilitas Kampus Universitas Dengan Menggunakan Klasifikasi *Support Vector Machine* (SVM)[20]	Universitas Telkom (*Tel-U)*	Risk Minimization (SRM) untuk menemukan hyperplane terbaik	dalam pemilihan perguruan tinggi sebagai sarana untuk proses memajukan kehidupan berbangsa dan bernegara perlu melakukan adanya peningkatan mutu dan kualitas layanan yang diberikan kepada mahasiswa. Kepuasan mahasiswa dianggap sebagai salah satu masalah	dapat membantu menyelesaikan permasalahan dalam menemukan analisis sentimen kepuasan terhadap layanan dan fasilitas kampus universitas *Tel-U* serta hasil pengujian dengan akurasi pada skenario terbaik (10000 data dengan kernel linear) sebesar 69.3%.	penelitian dari survey layanan dan fasilitas kampus, sedangkan untuk penelitian ini menggunakan objek penelitian pada agoda.com dan tiket.com

No	Judul	Objek	Metode	Masalah	Hasil	Perbedaan
				utama perguruan tinggi yang harus dipecahkan agar terciptanya perguruan tinggi yang mampu menduduki peringkat nasional maupun internasional. Layanan yang berpengaruh cukup besar dalam hal ini adalah layanan akademik. Tingkat kepuasan mahasiswa terhadap layanan		

No	Judul	Objek	Metode	Masalah	Hasil	Perbedaan
				berorientasi pada tenaga pendidik (dosen) sebagai pemberi jasa dan kualitas layanan dalam sarana dan prasarana kegiatan perkuliahan.		
6	Analisis Sentimen Transportasi Online Menggunakan Support Vector Machine Berbasis Particle Swarm Optimization[21]	Sentimen Pengguna Transportasi Online (*Twitter*)	SVM-PSO. SVM digunakan karena memilki hyperplane untuk memisahkan kelas positif dan negative dengan memaksimalkan margin,	Analisis sentimen diperlukan untuk mengetahui opini masyarakat tentang fenomena transportasi online dengan masalah seperti kriminalitas dan penipuan di Indonesia yang	Hasil penelitian ini menunjukkan bahwa penggunaan *SVM-PSO* lebih baik dengan nilai akurasi sebesar 96,04% daripada *SVM* biasa dengan nilai 95,46% meskipun	Perbedaan penelitian sebelumnya menggunakan objek penelitian dari data *tweet* pada *Twitter* @ambonlima transportasi online, sedangkan untuk penelitian ini menggunakan objek

No	Judul	Objek	Metode	Masalah	Hasil	Perbedaan
			sedangkan PSO digunakan untuk meningkatan akurasi	memicu pro dan kontra pada pengguna Twitter.	menggunakan parameter default.	penelitian pada agoda.com dan tiket.com
7	Analisis Sentimen Terhadap Layanan Indihome Berdasarkan Twitter Dengan Metode Klasifikasi Support Vector Machine (SVM)[22]	Sentimen Pengguna Layanan Indihome (*Twitter*)	SVM digunakan karena memiliki perhitungan masalah linear dengan menerapkan transformasi matematis menggunakan fungsi kernel	Indihome sebagai perusahaan Internet Service Provider (ISP) dengan keterbatasan ruang diskusi baik keluhan maupun kepuasan pengguna terhadap layanan indihome menjadikan Analisis ini diperlukan untuk	Hasil penelitian ini menunjukkan penggunaan SVM dapat membantu menyelesaikan permasalahan terhadap kepuasan penggunaan layanan Indihome dengan nilai *accuracy* 87%, *precision* 86%, *recall* 95%, *error rate* 13%, *f1-score* 90%, dan dari hasil	Perbedaan penelitian sebelumnya menggunakan data yang diambil melalui *Twitter*, sedangkan untuk penelitian ini menggunakan objek penelitian pada agoda.com dan tiket.com

No	Judul	Objek	Metode	Masalah	Hasil	Perbedaan
				mengetahui opini tentang pelayanan Indihome.Twitter sebagai sarana komunikasi dan informasi mengenai layanan indihome.	ulasan didapatkan 18,4% sentimen positif dan 81,6% sentimen negatif sehingga disimpulkan tingkat penggunaan layanan indihome cukup rendah	
8	Penerapan Algoritma Support Vector Machine Dalam Sentimen Analisis Terkait Kenaikan Tarif BPJS Kesehatan[23]	Sentimen Terkait kenaikan Tarif BPJS Kesehatan (*Twitter*)	SVM digunakan untuk menemukan hyperplane dan merupakan sebuha metode klasifikasi pada Machine Learning	Analisi sentimen ini digunakan untuk mengetahui opini tentang satu kebijakan pemerintah yang sedang ramai diperbincangan oleh netizen di	Hasil penelitian menunjukkan metode SVM dapat membantu menyelesaikan permasalahan dengan nilai accuracy 92% (87% sentimen positif dan	Perbedaan penelitian sebelumnya menggunakan data yang diambil melalui *Twitter*, sedangkan untuk penelitian ini menggunakan objek penelitian pada

No	Judul	Objek	Metode	Masalah	Hasil	Perbedaan
			(Supervised Learning) untuk memprediksi kelas menggunakan pola training	jejaring sosial twitter adalah kenaikan tarif BPJS kesehatan. Besarnya jumlah pengguna BPJS menyebabkan BPJS harus menyediakan layanan umpan balik kepada pengguna untuk mengetahui respon pengguna terhadap layanan BPJS.	96% sentimen negatif). Hasil klasifikasi dengan 671 sample data terdapat 271 data yang setuju dan 400 data yang tidak setuju terhadap kebijakan kenaikan Tarif BPJS Kesehatan.	agoda.com dan tiket.com

2.2 Dasar Teori

Adapun isi dari dasar teori adalah memuat teori – teori penting dalam penyusunan penelitian yang akan dilakukan oleh penulis, berikut dasar teori yang disusun oleh penulis dalam proses penelitian.

2.2.1Analisis Sentimen

Secara singkat analisis sentimen yaitu proses dalam memahami, mengekstrak dan mengolah data secara tekstual yang terkandung dalam suatu kalimat[18]. Analisis sentimen merupakan ilmu untuk mengenali dan mengekstrasi pendapat masyarakat dalam bentuk teks *Natural Language Processing (NLP)* yang biasanya dilakukan untuk mengetahui respon pengguna media sosial dalam memahami sebuah studi kasus yang membaginya menjadi tiga kelompok yaitu opini positif, negatif, dan netral [24].

Analisis sentimen atau dapat disebut dengan *opinion mining* merupakan analisa pendapat, sentimen, evaluasi tentang penilaian emosi seseorang terhadap topik tertentu yang kemudian diolah secara komputasi linguistik dan *text mining* [25]. Analisis ini biasanya digunakan untuk menganalisis penilaian opini, baik opini yang menunjukan ekspresi senang maupun tidak senang terhadap barang dan jasa. Informasi ini dapat bersifat *subjektif* yang terdiri dari nilai *positive* dan *negative* maupun *neutral* yang menjadikan nilai ini sebagai parameter [9].

2.2.2 Teks Mining

Text mining pertama kali diperkenalkan oleh *Fledman* dan *Ido Dagan* sebagai bentuk pengetahuan teks pada proses penggalian atau penambangan informasi teks berkualitas tinggi yang biasana teks data yang dimaksud yaitu data tidak terstruktur [26][27][28]. *Text mining* merupakan bagian dari data mining yang merupakan proses untuk mengklasifikasikan dokumen, data yang diperoleh memiliki pola yang benilai dari beperapa data yang berjumlah tinggi [29]. Tujuan dari text mining yaitu untuk menganalisis dokumen tesktual seperti email, ulasan, text biasa, halaman web, laporan dan dokumen resmi dengan mengekstrak data tersebut dan mengubahnya menjadi informasi yang

lebih berguna. Penambangan teks termasuk teknik linguistik, statistik pembelajaran mesin dan analisis dengan kata lain tujuan dari text mining yaitu ekstrasi pengetahuan pola dari berbagai dokumen teks untuk menganalisis topik dari media social dan mengklasifikasikannya kedalam analisis sentimen [30].

2.2.3 TD-IF

TF-IDF (Term Frequency - Inverse Document Frequency) merupakan statistik data numerik yang menunjukan pentingnya kata dalam dokumen yang sudah diklasifikasi, jika bernilai 1 maka kata tersebut terdapat dalam kalimat dan jika benilai 0 maka kata tersebut tidak ada dalam kalimat. TF-IDF biasanya digunakan untuk menghitung bobot dalam pengambilan informasi, semakin banyaknya kata yang muncul maka nilai TF-IDF semakin meningkat, dan apabila frekuensi kata sering muncul maka nilai TF-IDF semakin turun [31]. Berikut merupakan langkah-langkahnya dalam melakukan TF-IDF [32][33][34] :

$$TF\ (t,d) = 0.5 + 0.5\ \frac{F(t,d)}{\max\{f(w,d) : w \in d} \tag{2.1}$$

$$IDF\ (t,d) = log\frac{N}{Df(t,d)} \tag{2.2}$$

$$TF - IDF\ (t,d,D) = tf(t,d)xIDF(t,d) \tag{2.3}$$

Keterangan :

d = dokumen ke-d

t = kata ke-t dari kata kunci

W = bobot dokumen ke-d terhadap kata ke-t

TF = banyaknya kata

IDF = Inversed Dokumen Frequency

D = total dokumen

Df = banyaknya dokumen

Contoh perhitungan manual TF-IDF [35] :

a. Kalimat di setiap dokumen

Dokumen	Kalimat
D1	I loved this book
D2	This book is a absolutely good
D3	The best book ever

b. Bag of words

TF	i	loved	this	book	is	absolutely	good	best	book	ever
D1	1	1	1	1	0	0	0	0	0	0
D2	0	0	1	1	1	1	1	0	0	0
D3	0	0	0	1	0	0	0	1	1	1

c. Hasil perhitungan TF

TF	i	loved	this	book	is	absolutely	good	best	book	ever
D1	1/4	1/4	1/4	1/4	0	0	0	0	0	0
D2	0	0	1/5	1/5	1/5	1/5	1/5	0	0	0
D3	0	0	0	1/4	0	0	0	1/4	1/4	1/4

Keterangan :

1/4 = 0,25

1/5 = 0,2

d. Hasil perhitungan IDF

TF	i	loved	this	book	is	absolutely	good	best	book	ever
D1	Log (3/1)	Log (3/1)	Log (3/2)	Log (3/3)	0	0	0	0	0	0
D2	0	0	Log (3/2)	Log (3/3)	Log (3/1)	Log (3/1)	Log (3/1)	0	0	0
D3	0	0	0	Log (3/3)	0	0	0	Log (3/1)	Log (3/1)	Log (3/1)

Keterangan :

Log(3/1) = 0,477

Log(3/2) = 0,176

Log(3/3)=0

e. Hasil TF-IDF

TF	i	loved	this	book	is	absolutely	good	best	book	ever
D1	0,12	0,12	0,04	0	0	0	0	0	0	0
D2	0	0	0,03	0	0,09	0,09	0,09	0	0	0
D3	0	0	0	0	0	0	0	0,12	0,12	0,12

2.2.4 *Particle Swarm Optimization*

Particle Swarm Optimization (PSO) yang diperkenalkan oleh Kennedy dan Eberhart pada tahun 1995 merupakan metode optimasi sederhana untuk memodifikasi beberapa parameter. Optimasi ini dilakukan dengan menyeleksi atribut *(attribute selection)* dan *feature selection* yang dapat meningkatkan bobot atribut sehingga dapat meningkatkan akurasi untuk model *SVM*. Kelebihan dari *PSO* yaitu, sederhana, mudah diterapkan dan kecepatan dalam konvergensinya. Pencarian solusi menggunakan *algoritma PSO* berdasarkan partikel dalam populasi tertentu. Populasi ini dinilai secara acak dan memiliki batasan *minimum* dan batasan *maksimum*. Partikel ini melacak solusi dengan cara melalui ruang pencarian dengan cara mencari letak *terbaiknya (local best)* dan melacak letak partikel terbaik pada seluruh kelompok *(global best)* saat melalui search space [18][10][21]. Berikut merupakan langkah-langkah dalam menggunakan Particle Swarm Optimization (PSO) [36]:

a. Inisialisasi Partikel

Sebuah partikel untuk setiap dimensi yang berada dalam domain didefinisikan oleh dua vektor yaitu vektor minimal dan vektor maksimal (batas bawah dan batas atas).

$$x_i(0) = x_{\min d} + r_d(x_{\max d})\, Vd$$
$$= 1, \ldots n_x V = 1, \ldots n_s \qquad (2.4)$$

Keterangan :

x_{id} = posisi partikel ke-i pada dimensi d

r_d = range

V_{id}= Kecepatan setiap partikel, menjadi 0 yaitu, V_{id} (0) = 0

V_{id} merupakan kecepatan partikel ke-i pada dimensi ke-d. Posisi terbaik

b. Fungsi Objektif PSO

Fungsi objektif ini digunakan untuk mendapatkan nilai *fitness* dalam suatu partikel. Optimasi ini dapat memaksimalkan fungsi h, dimana $f = h$. Namun, untuk memperoleh nilai *minimal* fungsi h dapat menggunakan rumus.

$$f = {}^{1}/_{h}.$$

$$Fitness = x_{optimal} + \frac{1}{x\_min} \qquad (2.5)$$

Keterangan :

X_{min} = batas bawah dari x

X_{max} = batas atas dari x

c. Personal Best PSO

Personal Best PSO digunakan untuk membandingkan nilai fitness saat ini dan nilai fitness sebelumya.

$$pBest\ (t+1) = \begin{cases} pBest_i(t)\ Jika\ f(x_{id}(t+1)) \geq f pBest_i(t)) \\ x_{id}(t+1)\ Jika\ f(x_{id}(t+1)) < f(pBest_i(t)) \end{cases} \qquad (2.6)$$

Keterangan :

$pBest_i\ (t+1)$ = $pBest$ iterasi sekarang

$pBest_i\ (t)$ = $pBest$ iterasi sebelumnnya

$f(x_{id}(t+1))$ = nilai fitness iiterasi sekarang

d. Global Best PSO

Pencarian dengan mencari argument dari nilai maksimum pada semua $pBest$ dalam iterasi t.

$$k = argMax_{i=1}^{n}\{f(pBest_i(t))\} \qquad (2.7)$$

Keterangan :

$gBest(t) = pBest_k(t)$

$argMax_{i=1}^{n}$ = argument maksimal

$f(pBest_i(t))$ = nilai fitness $pBest_i$ iterasi (t)

e. Update kecepatan dan posisi

Kecepatan (V_{id}) dan posisi dari partikel (x_i) diubah.

$$V_{id}^{t+1} = w * v_{id}^{t} + c_i * Rand() * (p_{id} - x_{id}^{t}) + c_2 * Rand * p_{gd} - x_{id}^{t}) \quad (2.8)$$

Selanjutnya untuk membandingkan antara kecepatan dan v_{max} menggunakan persamaan 2.9

$$v_{max} = k * (v_{max} - v_{min}) \quad (2.9)$$

Persamaan perubahan posisi X_i :

$$x_{id}^{t+1} = x_{id}^{t+1} + v_{id}^{t+1} \quad (2.10)$$

Persamaan W inersia (w) :

$$w(i) = w_{max} - \left(\frac{w_{max} - w_{min}}{i_{max}}\right) x\ i \quad (2.11)$$

v_{max} dan v_{min} merupakan nilai awal dan nilai akhir untuk bobt inersia yaitu 0.9 dan 0,4. i_{max} merupakan jumlah iterasi maksimum.

Keterangan :

v_{id}^{t} = kecepatan partake ke-i untuk dimensi ke-d pada iterasi ke-t

x_{id}^{t} = posisi partikel ke-i untuk dimensi ke-d pada iterasi t

w = bobot inersia

c_1, c_2 = komponen kognitif dan *social*

p_{id} = $gBest$ atau posisi terbaik partikel pada x_{id}^{t}

p_{id} = $gBest$ atau seluruh posisi terbaik partikel

f. Kondisi berhenti

Jika sudah mencapai nilai iterasi maksimal atau perulangan telah mencapai nilai konvergen, maka perulangan berhenti dan nilai optimumnya didapatkan namum jika belum maka perulangan akan terus berlanjut.

Berikut merupakan contoh perhitungan manual *Particle Swarm Optimization* Pada studi kasus pemenuhan kebutuhan gizi balita [37]:

a. Inisialisasi Kecepatan Awal

Memiliki kecepatan dan kecepatan awal setiap partikel di set dengan nol.

b. Inisialisasi Partikel

Partikel	Nilai Dimensi					*Fitness*
	PH	PN	....	SA	SB	
X1	18	12		25	60	186.72
	11	5		60	3	
	60	46		49	33	
X2	46	30		59	33	190.5
	12	38		11	3	
	43	50		5	36	

c. Inisialisasi Pbest dan Gbest

Pada awal *Iterasi,* nilai *Pbest* disamakan dengan posisi awal partikel dan nilai Gbest didapatkan dari nilai fitness Pbest yang tertinggi.

Partikel	Nilai Dimensi					*Fitness*
$Pbest_1$	18	12	...	25	60	186.72
	11	5		60	3	
	60	46		49	33	
$Pbest_2$	46	30		59	33	190,5
	12	38		11	3	
	43	50		5	36	

Partikel	Nilai Dimensi					*Fitness*
$Gbest_1$	46	30	...	59	33	190,5
	12	38		11	3	
	43	50		5	36	

d. Update Kecepatan

Berikut ini contoh update kecepatan dan menghitung nilai w pada iterasi=1:

$W = 0.7 - \left(\frac{0.7-0.4}{2}\right) x\ 1 = 0{,}55$

$V_1^1 = WV_1^1 + c_1 rand_1\ x\ (Pbest_1 - x_1^1) +$

$c_2 rand_2\ x\ (Gbest_1 - x_1^1)$

$V_1^1 = 0.55 * 0 + 2 * 0.9342\ x\ (18 - 18) + 2$

$* 0.3963\ x\ (46 - 18) = 22.2$

e. Update Posisi dan Hitung Fitness

- menghitung update posisi

  $x_j^{k+1} = x_j^k + v_j^{k+1}$

  $x_1^{0+1} = x_1^0 + v_1^{0+1}$= 18 + 22 = 40

- Jika penalti dan total harga Partikel X1 yang didapat adalah 721.1 dan 40886.63. Penalti dan Total Harga Partikel X2 adalah 710.725 dan 50501.59. Sedangkan variasi Partikel X1 dan Partikel X2 adalah 26 dan 30, maka:

  $Fitness(x1,1) = \frac{1}{721.07}100000 + \frac{1}{721.07}100000 + 28$

  = 191.13

  $Fitness(x2,1) = \frac{1}{710.725}100000 + \frac{1}{50501.59}100000 + 30$

  = 195.50

- Setelah melakukan update posisi dapat melakukan normalisasi Posisi, digunakan agar perpindahan posisi tidak melebihi range yang telah ditentukan. Pada contoh perhitungan ini menggunakan batas bawah = 1 dan batas atas = 65, sehingga apabila ada perpindahan posisi yang melebihi batas atas maka di set menjadi 65 dan apabila kurang dari batas bawah maka posisi di set menjadi 1.

f. Update Pbest dan Gbest

Nilai *Pbest* baru baru didapatkan dengan cara membandingkan nilai *fitness* partikel baru dan *fitness Pbest* sebelumnya. Nilai *fitness* terbesar dijadikan sebagai Gbest terbaru.

2.2.5 *Support Vector Machine*

Support Vector Machine (SVM) diperkenalkan oleh Vladimir Vapnik pada tahun 1963 yang kemudian bekerja sama dengan Alexe Chernonekis untuk teori VC, keduanya sangat berjasa untuk SVM, Algoritma SVM sangat bermanfaat dan berhasil dalam penelitian salah satunya kategori teks[38]. *Support Vektor Machine (SVM)* merupakan salah satu algoritma yang menggunakan prinsip *Struktural Risk Minimmization* atau *SRM* [29][20].

Algoritma SVM merupakan algoritma yang sangat cepat dan efektif untuk masalah klasifikasi teks, hal ini dapat diketahui dari *hyperplane* di ruang fitur yang memsisahkan kelas *positve* dan kelas *negative* [26]. SVM memiliki konsep dan teori yang terstruktur dan baik, sehingga SVM merupakan metode dengan akurasi terbaik di bidang klassifikasi teks khususnya sentimen klasifikasi [39]. Algoritma ini memiliki konsep sederhana yaitu mencari hyperplane terbaik yang berfungsi untuk memisahkan dua kelas (positve dan negative) pada input space[17]. *Hyperplane* terbaik dapat ditemukan dengan cara mengukur *margin* dan mencari titik maksimalnya, *pattern* yang paling dekat disebut *support vector* [17].

a. *Hyperplane* merupakan garis atau batasan untuk memisahkan kelas. *Suuport vector* merupakan titik terdekat dari kedua kelas dan *margin* merupakan jarak antara *hyperplane* dan *support vector*, sehingga untuk mendapatkan *hyperplane* yang maksimal atau optimal dapat menggunakan *hyperplane* dengan cara memaksimalkan *margin*. Berikut persamaan untuk mencari nilai hyperplane berdimensi d :

$$\vec{w}.\vec{x} + b = 0 \qquad 2.6.1$$

w = nilai bobot
x = nilai atribut
b = nilai bias

Untuk pattern x_i termasuk class -1 dapat menggunakan persamaan

$$\vec{w}.\vec{x_i} + b \leq -1 \qquad 2.6.2$$

Untuk pattern x_i termasuk class +1 dapat menggunakan persamaan

$$\vec{w}.\vec{x_i} + b \geq +1 \qquad 2.6.3$$

Margin yang terbaik dapat ditemukan dengan cara memaksimalkan hyperplane dan titi terdekat yaitu, 1/||w|| . Mecari titik minimal atau disebut *Quadratic Programming* dapat diketahui dengan persamaan

$$\min_{\vec{w}} \tau(w) = \frac{1}{2}\|\vec{w}\|^2 \qquad 2.6.4$$

$$y_i(\vec{x_i}.\vec{w} + b) - 1 \geq 0, Vi \qquad 2.6.5$$

Lagrange Multiplier merupakan teknik komputasi yang digunakan untuk memecakan problem seperti ini, menggunakan persamaan

$$L(\vec{w}, b, a) = \frac{1}{2}\|\vec{w}\|^2 - \sum_{i=1}^{l} a_i \left(y_i\left((\vec{x_i}.\vec{w} + b) - 1\right)\right)$$

$$(i = 1,2, \dots, l) \qquad 2.6.6$$

α_i merupakan Lagrenge Multipliers bernilai nol atau positif ($\alpha \geq 0$), nilai optimal dapat dihitung dengan meminimalkan L terhadap w & b dan memaksimalkan L terhadap α_i . Titik optimal *gradient* $L = 0$ dapat dimodifikasi dengan memaksimalkan problem yang hanya menggandung α_i, seperti dibawah ini :

- Maximize

$$\sum_{i=1}^{l} a_i - \frac{1}{2}\sum_{i=1}^{l} a_i\, a_j y_i y_j \vec{x_i}.\vec{x_j} \qquad 2.6.7$$

- Subject to

$$a_i \geq 0 (i = 1,2, \dots, l) \sum_{i=1}^{l} a_i\, y_i = 0 \qquad 2.6.8$$

Data yang berkorelasi α_i dengan bernilai *positive* dapat disebut dengan *support vector*.

b. *Soft Margin* merupakan teknik yang bertujuan untuk mendapatkan *input space* pada dua buah kelas agar dapat terpisah secara sempurna, oleh karane itu persamaan 2.6.5 dapat dimodifikasi menjadi dengan menambahkan ε_i ($\varepsilon_i > 0$). Dapat diketahui dengan persamaan

$$y_i(\overrightarrow{x_i}.\vec{w} + b) \geq 1 - \xi_i, \qquad Vi \qquad 2.6.9$$

Dengan demikian persamaan 2.6.4 dapat diubah menjadi

$$\min_{\vec{w}} \tau(w, \xi) = \frac{1}{2}\|\vec{w}\|^2 + C\sum_{i=1}^{l} \xi_i \qquad 2.6.10$$

Parameter C ditambahkan berfungsi untuk mengontrol *tradeoff* antara *margin* dan *error* klassifikasi ε, maka dari itu nilai C yang besar berarti akan memberikan hasil yang lebih besar terhadap *error* klassifikasi.

c. Kernel Trick atau Kernel Linear

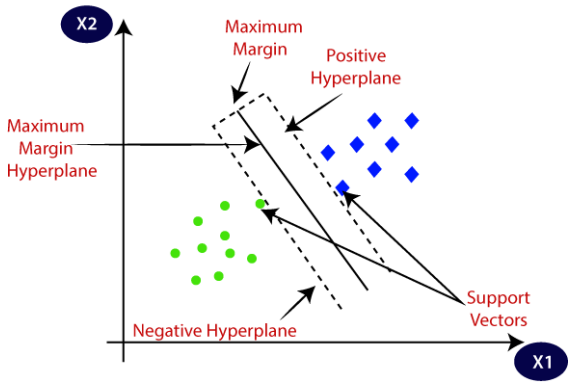
Kernel trick dapat memberikan berbagai manfaat seperti kemudahan proses pembelajaran SVM dan untuk menentukan support vector hanya menggunakan fungsi kernel yang digunakan. Berikut merupakan persamaan dalam mencari kernel trick

$$f\left(\Phi(\overrightarrow{x_i})\right) = \vec{w}.\Phi(\vec{x}) + b \qquad 2.6.13$$

$$= \sum_{i=1,\overrightarrow{x_i},\in SV}^{n} a_i\, y_i\, \Phi(\vec{x}).\Phi(\overrightarrow{x_i}) + b \qquad 2.6.14$$

$$= \sum_{i=1,\overrightarrow{x_i},\in SV}^{n} a_i\, y_i\, K(\vec{x}, \overrightarrow{x_i}) + b \qquad 2.6.15$$

SV pada persamaan diatas merupakan *subset data training set* yang terpilih sebagai *support vector* atau data x_i berkorepondensi pada $x_i \geq 0$. Berikut gambar 2.1 cara kerja algoritma *Support Vector Machine :*



Gambar 2.1 Cara kerja Algoritma *Support Vector Machine (SVM)*[23].

Contoh peritungan manual SVM :

a. Contoh studi kasus SVM

x_1	x_2	Kelas (y)
1	1	1
1	-1	-1
-1	1	-1
-1	-1	-1

$$y_i(w_1 * x_1 + w_2 * x_2 + b) \geq 1$$

b. Mencari persamaan

$y_i(w_1 * x_1 + w_2 * x_2 + b) \geq 1$

$1(w_1 * 1 + w_2 * 1 + b) \geq 1$

$(w_1 + w_2 + b) \geq 1$ (Persamaan 1)

$y_i(w_1 * x_1 + w_2 * x_2 + b) \geq 1$

$-1(w_1 * 1 + w_2 * -1 + b) \geq 1$

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

$y_i(w_1 * x_1 + w_2 * x_2 + b) \geq 1$

$-1(w_1 * -1 + w_2 * 1 + b) \geq 1$

$(w_1 - w_2 - b) \geq 1$ (Persamaan 3)

$y_i(w_1 * x_1 + w_2 * x_2 + b) \geq 1$

$-1(w_1 * -1 + w_2 * -1 + b) \geq 1$

$(w_1 + w_2 - b) \geq 1$ (Persamaan 4)

c. Tahap eliminasi

Menjumlahkan persamaan (1) dan persamaan (2) :

$(w_1 + w_2 + b) \geq 1$

$(-w_1 + w_2 - b) \geq 1$

---------- +

$2w_2 = 2$

$w_2 = 1$

Mejumlahkan persamaan (1) dan persamaan (3) :

$(w_1 + w_2 + b) \geq 1$

$(w_1 - w_2 - b) \geq 1$

---------- +

$2w_1 = 2$

$w_1 = 1$

Menjumlahkan persamaan (2) dan persamaan (3)

$(-w_1 + w_2 - b) \geq 1$

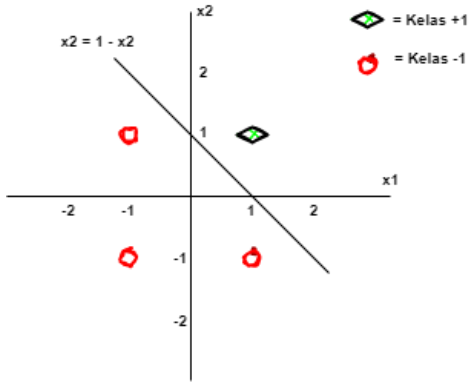
$(w_1 - w_2 - b) \geq 1$

---------- +

-2b =2

b = -1

Berikut merupakan gambar 2.2 Hyperplane :



Gambar 2.2 *Hyperplane*[38]

2.2.6 *Confusion Matrix*

Confusion Matrix merupakan pengukuran atau perhitungan dari hasil yang didapatkan[20]. Berikut merupakan persamaannya.

a. *Precision* merupakan proporsi kasus dengan positif benar

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

b. *Recall* merupakan proporsi positif yang teridentifikasi benar

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

c. *Accuracy* merupakan pebandingan kasus identifikasi benar dengan jumlah semua kasus

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

TP (True Positive) yaitu jumlah dokumen komentar positif diprediksi positif oleh sistem, *FN (False Negative)* yaitu jumlah dokumen komentar positif diprediksi negatif oleh sistem, *FP (False Positive)* yaitu jumlah dokumen komentar negatif di prediksi positif oleh sistem dan TN *(True Negative)* yaitu jumlah dokumen komentar negative di prediksi negatif oleh sistem.

2.2.7 *K-Fold Cross Validation*

Cross validation merupakan pengukuran untuk menilai kinerja model prediktif dan analisis statistik, salah satunya K-fold cross validation yang merupakan teknik membagi sampel asli secara acak menjadi K sub sampel kemudian satu sub-sampel sebagai data validasi untuk menguji model dan sub-sampel K-1 yang tersisa sebagai pelatihan. Berikut ini merupakan langkah-langkahnya[40]:

1. Bagi data K secara kasar menjadi bagian yang sama
2. i = 1,2,3... K, sesuaikan model dengan parameter y atau K-1 lainya, berikan

   $\alpha^{-k}(y)$, hitung kesalahanya dalam memprediiksi bagian ke-K

   $$E_k(y) = \sum_{i\epsilon\ kth\ part}[y_{i+} + x_i\alpha^{-k}(y)^2] \qquad 2.6.16$$

3. Lakukan secara berulang, hingga menemukan nilai dengan kesalahan terkecil