BAB II

TINJAUAN PUSTAKA DAN LANDASAN TEORI

2.1 Tinjauan Pustaka

Penelitian ini dilakukan dengan mencari sumber referensi dan informasi dari jurnal penelitian yang relevan dengan penelitian yang akan dilakukan oleh penulis. Kajian yang digunakan dapat dijadikan perbandingan antara penelitian terdahulu dan penelitian yang hendak dilakukan baik dari segi perbedaan dan hasil yang diperoleh.

Penelitian oleh Abdulloh dan Pambudi pada tahun 2021 berjudul "Analisis Sentimen Pengguna Youtobe Terhadap Program Vaksin Covid-19". Penelitian ini melakukan klasifikasi terhadap pro dan kontra dari komentar masyarakat pada media sosial Youtube terhadap vaksinasi. Dalam pengolahan opini masyarakat pada penelitian ini menggunakan metode *Support Vector Machine* dengan tujuan sebagai alternatif dalam menentukan respon publik terhadap suatu peristiwa, dari pengolahan teks dapat dibuat model klasifikasi untuk dapat dijadikan informasi serta masukan kepada pihak tertentu sebagai bahan pertimbangan. Hasil kesimpulan yang diperoleh menyebutkan bahwa *Support Vector Machine* terbukti baik dalam melakukan klasifikasi teks terutama klasifikasi komentar positif dan negatif pada platform Youtobe[13].

Penelitian oleh Utama dan kawan-kawan pada tahun 2019 berjudul "Analisis Sentimen Sistem Ganjil Genap di Tol Bekasi Menggunakan Algoritma *Support Vector Machine*". Pada penelitian tersebut melakukan analisis sentimen terhadap efektifitas sistem ganjil dan genap tol bekasi pada komentar media sosial Twitter, Instagram, Youtube, dan Facebook menggunakan model SVM dari hasil *Confussion Matrix* mendapatkan nilai *accuracy* sebesar 78,18%, *Precision* sebesar 74,03%, atau *Recall* sebesar 86,82%. Dengan kesimpulan yang di dapat bahwa penggunaan Algoritma *Support Vector Machine* dapat menganalisis sentimen ganjil genap di tol bekasi[1].

Penelitian oleh Asrianto dan Herwinanda pada tahun 2022 berjudul "Analisis Sentimen Kenaikan Harga Kebutuhan Pokok di Media Sosial Youtube Menggunakan Algoritma *Support Vector Machine*" melakukan pengklasifikasian terhadap sentimen positif dan sentimen negatif menggunakan empat model klasifikasi SVM dan hasilnya menunjukkan bahwa SVM linear dengan SMOTE mendapatkan tingkat akurasi terbaik dengan nilai akurasi tertinggi sebesar 86,33%, diikuti oleh presisi 75%, *f1-score* 70,59%. Evaluasi performa akurasi pada SVM linear dengan hasil hitung data validasi diperoleh nilai akurasi sebesar 72%[14].

Penelitian oleh Fahlevvi pada tahun 2022 yang berjudul "Analisis Sentimen Terhadap Ulasan Aplikasi Pejabat Pengelola Informasi dan Dokumentasi Kementerian Dalam Negeri Republik Indonesia di *Google Playstore* Menggunakan Metode *Support Vector Machine*". Dalam penelitian ini melakukan analisis sentimen berdasarkan ulasan pengguna pada aplikasi PPID di *Google Play Store* menggunakan metode *Support Vector Machine* dan TF-IDF dalam pembobotan karakter. Hasil penelitian menunjukkan bahwa analisis menggunakan SVM menghasilkan *k-fold* 88%, *precision* 94%, *recall* 100%, *f-measure* 97%, dan *accuracy* sebesar 97% dan menyatakan bahwa metode SVM dapat berjalan dengan baik dalam melakukan analisis sentimen[15].

Penelitian oleh Amalia dan kawan-kawan pada tahun 2021 yang berjudul "Analisis Sentimen Review Pelanggan Restoran Menggunakan Algoritma *Support Vector Machine* dan *K-Nearest Neighbor*" penelitian melakukan analisis *review* pelanggan restoran menggunakan 2 kelas yaitu positif dan negatif dengan metode klasifikasi *Support Vector Machine* dan *k- Nearest Neighbor* dan menggunakan metode Crisp-dm untuk membandingkan hasil klasifikasi dari kedua algoritma. Hasilnya menunjukkan bahwa SVM memiliki hasil kinerja baik daripada algoritma k-NN dengan nilai akurasinya sebesar 81,92% dan nilai AUC sebesar 0,918 sedangkan pada k-NN menghasilkan nilai akurasi sebesar 59,03% dan AUC sebesar 0,590[16].

Penelitian oleh Ratino dan kawan-kawan pada tahun 2020 berjudul "Sentimen Analisis Informasi Covid-19 Menggunakan *Support Vector Machine* dan *Naïve Bayes*" melakukan terhadap komentar sosial media Instagram terhadap

informasi Covid-19 dengan membandingkan dua algoritma yaitu SVM dan *Naïve Bayes* dengan optimasi dengan operator *Particle Swarm Optimization*. Hasilnya menunjukkan bahwa algoritma SVM berbasis PSO atau tidak menggunakan PSO, selalu mendapatkan akurasi lebih tinggi dari *Naïve Bayes* dengan selisih akurasi 2,21% dimana SVM sebesar 80,23% dan *Naïve Bayes* sebesar 78,02%. Sedangkan apabila menggunakan operator *Particle Swarm Optimazation* SVM menghasilkan akurasi sebesar 81,16 % dan *Naïve Bayes* sebesar 79,07%[10].

Penelitian oleh Rai B dan Shetty pada tahun 2019 dengan judul "*Sentiment Analysis Using Machine Learning Classifiers: Evaluation Of Performance*" melakukan penelitian mengenai analisis sentimen pendapat pada media sosial media twitter dengan polaritas positif, negatif, dan netral untuk label tweet menggunakan metode *machine learning* untuk melakukan perbandingan tiga klasifikasi yaitu algoritma *Random Forest*, *Naïve Bayes*, dan *Support Vector Machine* serta mengukur evaluasi kinerja dari metode tersebut seperti akurasi, presisi, dan *recall*. Hasilnya menyimpulkan bahwa ketiga algoritma bisa mendapatkan hasil yang baik apabila semakin banyak jumlah tweet maka akurasi meningkat dalam setiap kasus, jumlah tweet yang semakin banyak juga dapat membuat presisi meningkat secara proporsional[17].

Penelitian oleh Mehta dan Deshmukh tahun 2022 berjudul "*Youtube Ad View Setiment Analysis using Deep Learning and Machine Learning*" menganalisis sentimen tanggapan publik pada iklan di sosial media Youtube dengan algoritma *machine learning* dan *deep learning* seperti *Linear Regression*, *SVM* dengan pembobotan TF-IDF. Hasil penelitian menunjukkan bahwa SVM dengan kernel linier mendapatkan nilai terbaik dengan akurasi dan F1-score tertinggi, RBF SVM mendapatkan skor presisi tertinggi. RMSE tertinggi diperoleh oleh LR dan SVM masing-masing sebesar289.078 dan 288.736. Sedangkan nilai minimum diperoleh algoritma DT sebesar 260.193. Dan pembobotan kata Delta TF-IDF kombinasi SVM menghasilkan kinerja yang baik dalam melakukan analisis sentimen[20].

Tabel 2.1 Kajian Penelitian Terdahulu

No	Judul Penelitian	Peneliti	Objek Penelitian	Metode Penelitian	Hasil	Perbedaan
1	Analisis Sentimen Pengguna Youtube Terhadap Program Vaksin Covid-19 (2021)	Ferian Fauzi Abdulloh, Iqbal Rilo Pambudi	Melakukan analisis sentimen terhadap komentar pengguna Youtube mengenai program vaksin covid-19	SVM	Hasil pengujian sentimen menunjukkan sentiment negative sebesar 41% dan sentiment positif 17%. Support Vector Machine memiliki performa model yang baik pada kernel sigmoid dengan rata-rata akurasinya 77.75%. SVM dalam hal ini dapat mengklasifikasikan teks terutama dalam mengklasifikasi komentar negatif dan positif pada platform Youtube.	Dalam penelitian yang akan dilakukan pengujian dibagi menjadi 3 skenario yaitu 70:30, 80:20, 75:25 dan tahapan pengujian dilakukan satu kali pada 2 kelas sentiment positif dan negative, sedangkan dalam jurnal acuan pengujian dilakukan 1 kali skenario yaitu 70:30, dan uji SVM dilakukan dalam dua tahap pengujian, tahap pertama pada 3 kelas (positif, netral, negatif) dan dua kelas (positif,negatif) sentiment pada pengujian yang ke dua.

No	Judul Penelitian	Peneliti	Objek Penelitian	Metode Penelitian	Hasil	Perbedaan
2	Analisis Sentimen Sistem Ganjil Genap di Tol Bekasi Menggunakan Algoritma Support Vector Machine (2019)	Heru Sukma Utama, Didi Rosiyadi, Bobby Suryo Prakoso, Dedi Ariadarma	Mengetahui informasi sentimen masyarakat terhadap efektifitas sistem ganjil genap tol bekasi di media sosial	SVM	Support Vector Machine dapat menganalisis sentimen pada penelitian dengan *hasil confussion matrix*, yaitu akurasi sebesar 78,18%, presisi 74,03%, dan*recall* atau sensitivitas sebesar 86,82%.	Subjek dalam penelitian ini hanya 1 platform yang digunakan yaitu Youtube, sedangkan dalam jurnal menggunakan 4 subjek penelitian yaitu Twitter, Youtube, Facebook dan Instagram.
3	Analisis Sentimen Kenaikan Harga Kebutuhan Pokok di Media Sosial Youtube	Rudy Asrianto, Melda Herwinanda	Mengetahui opini masyarakat terhadap kenaikan harga kebutuhan	SVM	Hasil empat pengujian klasifikasi SVM didapatkan bahwa SVM linear dengan SMOTE memiliki keakuratan terbaik dengan nilai akurasi tertinggi sebesar 86,33%, presisi 75%, *recall* 66,67%	Dalam penelitian ini tidak menggunakan teknik penyeimbangan kelas data mengenai komentar positif dan negatif sedangkan dalam jurnal menggunakan Teknik SMOTE untuk mengatasi masalah *class*

No	Judul Penelitian	Peneliti	Objek Penelitian	Metode Penelitian	Hasil	Perbedaan
	Menggunakan Agoritma Support Vector Machine (2022)		pokok pada forum diskusi channel sosial media Youtube		dan nilai f1-*score* 70,59%. Pengujian pada data validasi (data uji baru diluar dataset) mendapat nilai akurasi 72%.	*imbalance problem* (CIP) yang bekerja dengan memodifikasi dataset yang tidak seimbang dengan cara membuat data sintetik baru dari kelas minoritas dengan tujuan meningkatkan kinerja dari metode klasifikasi.
4	Analisis Sentimen Terhadap Ulasan Aplikasi Pejabat Pengelola Informasi dan Dokumentasi Kementerian	Mohamm ad Rezza Fahlevvsi	Mengetahui sentimen megenai ulasan pengguna aplikasi PPID Kemendagri di Google Play Store	SVM	Penggunaan SVM dengan pembobotan TF-IDF pada ulasan pengguna aplikasi PPID Kemendagri di *Playstore* diimplemntasikan baik dengan rata-rata *k-fold* 88%, presisi 94%, *recall* 100%, *f-measure* 97% dan akurasi 97%.	Dalam jurnal ini menggunakan subjek *google play store*. Sedangkan penelitian menggunakan subjek Youtube. Output penelitian yang dilakukan berupa evaluasi model *confussion matrix*. Sedangkan pada jurnal outputnya tidak hanya

No	Judul Penelitian	Peneliti	Objek Penelitian	Metode Penelitian	Hasil	Perbedaan
	Dalam Negeri Republik Indonesia di Google Playstore Mengguakan Metode Support Vector Machine (2022)					*confussion matrix* tetapi juga *k-fold cross validation.*
5	Analisis Sentimen Review Pelanggan Restoran Menggunakan Algoritma Support Vector	Bintang Sifa Amalia, Yuyun Umaidah, Rini Mayasari	Melakukan sentimen analisis pada layanan pesan antar online pada restoran solaria dari	SVM dan KNN	Algoritma SVM mempunyai kinerja lebih baik dari algoritma K-NN dengan nilai akurasi sebesar 81,92%	Dalam jurnal yang dikaji menggunakan 2 metode yaitu SVM dan KNN untuk membandingkan performa kinerja yang lebih baik dari algoritma diantara keduanya. Sedangkan pada penelitian yang dilakukan hanya memuat

No	Judul Penelitian	Peneliti	Objek Penelitian	Metode Penelitian	Hasil	Perbedaan
	Machine dan K- Nearest Neighbor (2021)		review pelanggan			satu metode saja yaitu SVM dengan menggunakan pembobotan TF-IDF dan fitur *sentiwordnet* untuk membantu mengetahui skor sentiment. Objek dalam penelitian ini adalah data komentar Youtube sedangkan dalam jurnal data tweet pada twitter.
6	Sentimen Analisis Informasi Covid-19 menggunakan Support Vector Machine dan	Ratino, Noor Hafidz, Sita Anggraeni, Windu Gata	Mengetahui sentimen masyarakat melalui komentar pada media sosial instagram terhadap	SVM dan NB	NB memiliki akurasi sebesar 79,07% dan AUC 0,729, sedangkan SVM memiliki akurasi sebesar 81,16% dan AUC 0,903. Pada algoritma SVM baik berbasis PSO (*Particle Swarm ptimization*) ataupun tidak selalu	Pada jurnal yang dikaji penelitian dilakukan menggunakan dua algoritma yaitu SVM dan NB untuk melakukan perbandingan akurasi klasifikasi menggunakan optimasi dengan operator *Particle Swarm*

No	Judul Penelitian	Peneliti	Objek Penelitian	Metode Penelitian	Hasil	Perbedaan
	Naïve Bayes (2020)		informasi Covid-19		menghasilkan akurasi yang lebih tinggi.	*Optimization* (PSO). Sedangkan dalam penelitian yang akan dilakukan menggunakan satu metode yaitu SVM dan pembobotan TF-IDF serta pelabelan *sentiwordnet*.
7	*Sentiment Analysis Using Machine Learning Classifiers: Evaluation of Performance* (2019)	Shamantha Rai B, Sweekriti M Shetty	Mengevaluasi popularitas tweet sebagai positif dan negatif dari tweet atau ulasan yang diunggah di twitter	RF, SVM, NB	Semakin banyak jumlah tweet maka akurasi meningkat dalam setiap kasus, jumlah tweet yang semakin banyak presisi meningkat secara proporsional.	Pada penelitian yang dilakukan mengambil subjek dari youtube dan hanya menggunakan satu metode klasifikasi yaitu SVM sedangkan dalam jurnal ini menggunakan subjek Twitter dan menggunakan tiga mesin pengklasifikasian yaitu Random Forest, SVM dan Naïve Bayes

No	Judul Penelitian	Peneliti	Objek Penelitian	Metode Penelitian	Hasil	Perbedaan
8	*Youtube Ad View Setiment Analysis using Deep Learning and Machine Learning* (2022)	Tanvi Mehta, Ganesh Deshmukh	Melakukan analisis prediksi sentimen reaksi publik terhadap tayangan iklan pada Youtube untuk ekstrasi informasi dan visualisasi data	LR, SVM, DT, RF, ANN	RMSE tertinggi diperoleh oleh LR dan SVM masing-masing sebesar 289.078 dan 288.736. sedangkan nilai minimum diperoleh algoritma DT sebesar 260.193	Pada penelitian ini terdapat tiga skenario data testing dengan 3 skenario yaitu 70:30,80:20 dan 75:25. Untuk training model hanya menggunakan SVM berbeda dengan jurnal ini menggunakan satu skenario testing yaitu 80:20 dan menggunakan beberapa training model diantaranya *Algorithms like Linear Regression* (LR), *Support Vector Machine* (SVM), *Decision Tre*e (DT), *Random Forest* (RF), dan *Artificial Neural Network* (ANN).

No	Judul Penelitian	Peneliti	Objek Penelitian	Metode Penelitian	Hasil	Perbedaan
9	*Sentiment Analysis on Youtube social media Using Decision Tree and Random Forest Algorithm: A Case Study* (2020)	Mohammad Aufar, Rachmadita Andreswari, Dita Pramesti	Mengetahui sentimen mengenai komentar masyarakat dari konten Youtube terhadap produk Nokia	DT dan RF	DT memiliki akurasi yang lebih tinggi dibanding RF dengan perbedaan sedikit akurasi sebesar 89,4% dan RF sebesar 88,2%.	Dalam penelitian ini menggunakan pelabelan *sentiwordnet* dalam Analisa skor sentiment sedangkan pada jurnal ini pelabelan menggunakan VADER (*Valence Aware Dictionary and Sentiment Reasoner*) adalah alat analisis sentimen berbasis leksikon dan aturan yang secara khusus diatur untuk sentimen yang dinyatakan di media sosial
10	*Sentiment Analysis of Indonesian Movie Trailer on YouTube*	Muhammad Alkaff, Andreyan Rizky Baskara,	Mengetahui analisis sentimen terhadap komentar trailer	Delta TF-IDF, Regresi Logistik, NB dan SVM	*Regresi Logistik* dan NB baik dalam identifikasi sentimen untuk hal tertentu secara alir, sedangkan SVM mampu memberikan kinerja yang baik	Pada penelitian ini mengunakan metode pembobotan TF-IDF biasa, hanya bobot kata yang diperhitungkan, dan kata-kata positif dan negatif dianggap sama

No	Judul Penelitian	Peneliti	Objek Penelitian	Metode Penelitian	Hasil	Perbedaan
	Using Delta TF-IDF and SVM (2020)	Yohanes Hendro Wicaksono	film Indonesia pada Youtube		pada analisis sentimen untuk genre film secara umum	penting. Namun, dalam jurnal ini menggunakan, Delta TF-IDF untuk menghitung perbedaan (delta) dalam bobot kata IDF (Inverse Document Frequency) berdasarkan kehadiran kata-kata dalam dokumen positif dan negatif.

Pada pelaksanaan KTT G20 yang dilaksanakan di Bali menuai berbagai pro dan kontra dari masyarakat Indonesia, salah satunya ada pada komentar di siaran video yang di upload oleh channel sekretariat presiden pada sosial media Youtube. Dari pro dan kontra yang ada pada kolom komentar tersebut dapat dilakukan analisis sentimen. Dikarenakan Youtube menjadi salah satu media sosial teratas yang sedang banyak digemari oleh kalangan masyarakat dan memiliki kolom komentar yang tidak terbatas sehingga dapat memudahkan untuk mencari data yang akan digunakan dalam analisis sentimen. Berdasarkan beberapa pemaparan berbagai penelitian di atas, penelitian akan dilakukan menggunakan metode *Support Vector Machine* karena hasil akurasinya memiliki nilai yang tinggi dan performa cukup baik untuk mengukur analisis sentimen. Dalam melakukan penelitian ini mengacu pada penelitian dengan judul "Analisis Sentimen Pengguna Youtube Terhadap Program Vaksin Covid-19" oleh F. F. Abdulloh pada Tabel 2.1 Kajian Pustaka poin pertama karena pada penelitian tersebut menggunakan algoritma yang sama yaitu *Support Vector Machine* dan datanya diperoleh dari kolom komentar pada Youtube, dan pengujian data menggunakan 4 kernel SVM yaitu linear, polynomial, sigmoid, RBF untuk melihat kernel mana yang menghasilkan akurasi paling baik. Perbandingan penelitian sebelumnya terletak pada data yang diambil yaitu komentar Youtube mengenai program vaksin Covid-19 sedangkan yang akan dilakukan data dari komentar Youtube mengenai KTT G20. Selain itu perbedaan lain pada penelitian sebelumnya adalah pada data latih dan data uji yang digunakan yaitu 70:30 dengan pembagian tiga kelas sentimen yaitu positif, negatif, dan netral [13]. Sedangkan pada penelitian ini data latih dan data uji akan menggunakan 3 skenario yaitu 70:30, 75:25, 80:20 dengan pembagian 2 kelas sentimen positif dan negatif. Penelitian ini juga berfokus pada percobaan metode kernel pada algoritma SVM. Hal tersebut dilakukan untuk dapat mengetahui metode kernel terbaik pada algoritma SVM yang dapat melakukan klasifikasi sentiment dengan lebih akurat.

2.2 Dasar Teori

2.2.1 *Data Mining*

Data mining atau penambangan data adalah tindakan mengidentifikasi dan mendapatkan akses ke sejumlah data yang besar dan informasi yang belum ditemukan sebelumnya yang dapat dipahami, membantu mengisi basis data besar, dan dapat digunakan untuk mengembangkan keputusan bisnis. Tujuan dari *data mining* adalah untuk mendeskripsikan metode untuk mengidentifikasi pola tersembunyi. Untuk mengekstraksi dan mengidentifikasi informasi, proses penambangan data ini menggunakan metode statistik, aljabar, kecerdasan buatan, dan pembelajaran mesin. Proses KDD (*Knowledge Discovery in Databases*), yang mencakup langkah-langkah seperti pemilihan data, pra-pemrosesan, transformasi, penambangan data, dan evaluasi hasil, termasuk penambangan data sebagai salah satu komponennya[21].

2.2.2 API

Application Programming Interface (API) adalah sekumpulan instruksi dan pedoman yang membantu pemrogram membuat perangkat lunak untuk sistem operasi tertentu. Salah satunya adalah Youtube yang menawarkan dua jenis panggilan, yaitu REST dan XML-RPC, bagi developer untuk mendapatkan statistik video dan data dari channel Youtube. Pada penelitian ini, API yang digunakan untuk mengakses data komentar dari video pada Youtube.

2.2.3 Analisis Sentimen

Analisis sentimen adalah metode untuk mengenali, mengekstraksi, dan menggunakan informasi tekstual secara otomatis untuk mengidentifikasi informasi sentimental dalam gagasan. Analisis sentimen digunakan untuk menentukan apakah ide seseorang atau pendapat yang mirip dengan orang lain tentang subjek atau objek disukai atau tidak disukai[22].

2.2.4 *Text Preprocessing*

Data komentar dari Youtube merupakan data mentah yang harus diolah untuk meningkatkan konsistensi dan mengurangi variasi data yang ada melalui proses

data preprocessing yaitu data yang diperoleh dan disiapkan untuk dianalisis selanjutnya. Setelah dilakukan *data cleansing* pada data dari komentar Youtube. Kemudian proses selanjutnya yang dilakukan adalah mengembalikan sejumlah teks ke teks alami dengan meminimalkan *noise* pada tahap selanjutnya. Diawali memecah komentar menjadi bagian-bagian kata tertentu pada tokenisasi, *case folding* untuk merubah menjadi huruf kecil, menghilangkan *stopwords*, seluruhan kata yang terdapat imbuhan diubah menjadi kata dasar pada proses *stemming*[23].

2.2.5 TF-IDF

Metode *Term Frequency-Inverse Document Frequency* (TF-IDF) digunakan untuk membobot data. Sekelompok istilah atau kata akan divektorisasi atau diubah menjadi bentuk numerik selama prosedur ini. Ini penting karena hanya dapat menganalisis data dalam bentuk numerik untuk klasifikasi. Metode pembobotan ini merupakan gabungan antara *term frequency* dan *inverse document frequency*. Kuantitas *term* muncul dalam dokumen dikenal sebagai *term frequency*. Pembobotan yang diterapkan berhubungan langsung atau berbanding lurus dengan jumlah *term* yang muncul. *Inverse document frequency*, di sisi lain, adalah metode untuk menentukan seberapa signifikan kata-kata dalam sebuah dokumen. Nilai TF-IDF diperoleh menggunakan persamaan sebagai berikut[24].

Menghitung TF (*Term Frequency*)

$$TF = \begin{cases} 1 + log_{10}\,(f_{t,d}), & f_{t,d} > 0 \\ 0, & f_{t,d} = 0 \end{cases} \quad (2.1)$$

Dimana nilai ft,d adalah frekuensi term (t) pada document (d). misal suatu kata atau term terdapat dalam suatu dokumen sebanyak 5 kali maka diperoleh bobot = 1 + log (5) = 1.699. Tetapi jika term tidak terdapat dalam dokumen tersebut, bobotnya adalah nol (0).

Menghitung IDF (*Inverse Document Frequency*)

$$IDF = \log 10 \frac{N}{dft} \quad (2.2)$$

Menghitung TF-IDF

$$W_{t,d} = TF \,.\, IDF \tag{2.3}$$

Keterangan:

TF = Bobot kata setiap dokumen.

ft,d = Jumlah kemunculan term pada dokumen.

IDF = Bobot inverse dalam dokumen DF.

dft = Jumlah dokumen yang mengandung term.

Wt,d = Perhitungan pembobotan TF-IDF.

2.2.6 Sastrawi

Stemmer sastrawi adalah perpustakaan langsung dengan antarmuka yang intuitif. Kesulitan-kesulitan *stemming* yang dapat diatasi oleh algoritma-algoritma di *library* ini antara lain menghilangkan kata jamak, menghindari *understemming* dengan aturan tambahan, dan mencegah *overstemming* menggunakan kamus[25]. Dalam melakukan *stemming* dalam bahasa Indonesia dapat menggunakan *library* python sastrawi. Sastrawi merupakan *library* yang dapat mengubah kata berimbuhan dalam bahasa Indonesia menjadi bentuk kata dasar[26].

2.2.7 *SentiWordNet*

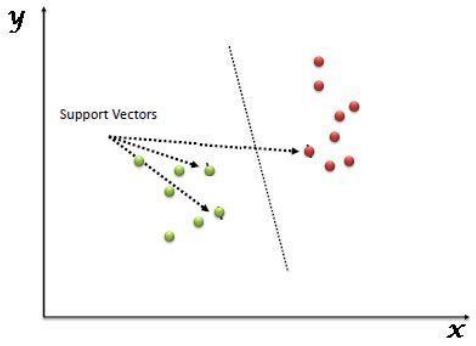
SentiWordNet (SWN) dalam menunjukkan popularitas (positif, negatif atau netral). SWN memiliki tiga skor numerik, skor tersebut mendefinisikan dari hasil dari setiap kata mengandung tingkat positif dan tingkat negatif atau netral. Untuk skor dengan kisaran 0,0 sampai 1,0 dan berjumlah 1,0 untuk setiap satu atau lebih sinonim. SWN memiliki sistem *random walk*. Apabila ada 2 konteks kata yang memiliki sinonim yang sama sentimennya akan sama. Sinonim atau disebut synset itu dikaitkan dengan konteks positif atau negatif. Jika konteks positif yang dihasilkan banyak yang berhubungan maka nilai positif yang dihasilkan pun akan semain besar dan sebaliknya[27]. Keefektifan *SentiWordNet* menunjukkan bagaimana memberikan skor sentiment pada kata dapat membantu meningkatkan akurasi dalam klasifikasi[28].

2.2.8 *Support Vector Machine* (SVM)

Klasifikasi merupakan salah satu teknik dalam data mining. Klasifikasi dilakukan untuk mengelompokkan objek yang memiliki karakteristik atau ciri yang sama ke dalam beberapa kelas. Biasanya klasifikasi dilakukan dalam menentukan ciri-ciri dengan diwakili oleh kalimat tertentu. Dalam menyelesaikan permasalahan dapat dengan mudah diselesaikan dengan penggunaan metode atau teknik[29]. Untuk mengklasifikasi data dapat digunakan metode algoritma klasifikasi seperti *Support Vector Machine*, *Naïve Bayes*, *Random Forest*, *Decision Tree*, CNN dan lain sebagainya.

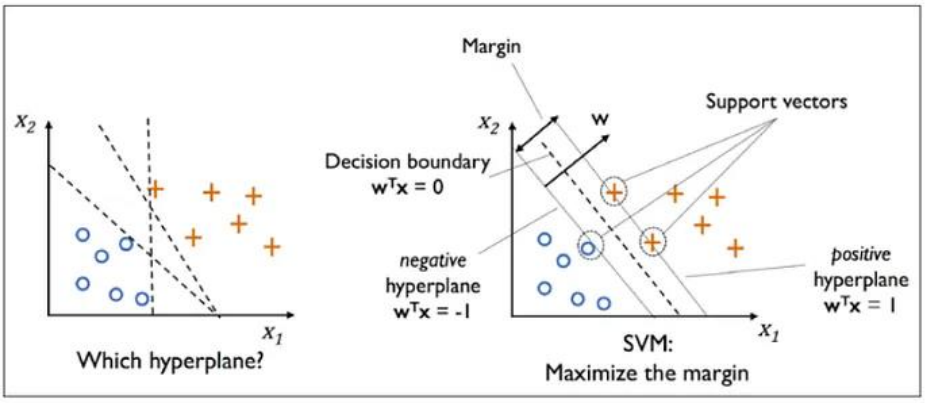
Support Vector Machine (SVM) adalah salah satu algoritma pembelajaran mesin yang tergolong dalam superviced learning[9]. SVM melakukan hasil dari proses pelatihan sebagai prediksi kelas. SVM merupakan teknik yang dapat melakukan prediksi baik dalam kasus klasifikasi maupun regresi. SVM melakukan pengelompokkan data untuk menemukan garis pemisah maksimal pada semua kelas yang merupakan permukaan sebuah keputusan yang disebut dengan *hyperlane atau Maximum Marginal Hyperplane* (MMH). SVM adalah metode pembelajaran mesin dengan tujuan menemukan *hyperplane* terbaik untuk mengklasifikasikan dua kelas yang disebut *binary classification* dan lebih dari dua kelas disebut *multi class classification*. Ini beroperasi pada konsep *Structural Risk Minimization* (SRM)[35]. Teori ini SVM dimulai ketika akan mengelompokan kasus-kasus linier yang dipisahkan dengan *hyperplane* dan dibagi menurut kelasnya.

Gambar 2.1 adalah visualisasi klasifikasi SVM dengan mencari *hyperplane* yang membedakan kedua kelas tersebut:



Gambar 2. 1 Visualisasi klasifikasi SVM

SVM diawali dengan masalah klasifikasi dua kelas sehingga membutuhkan set pelatihan positif dan negatif. *Margin* yaitu jarak terpendek dari suatu *hyperplane* dengan sisi *margin* lainnya dimana posisi kedua *margin* paralel dengan *hyperplane.* SVM akan berusaha mendapatkan *hyperplane* (pemisah) sebaik mungkin untuk memisahkan kedua kelas dan memaksimalkan *margin* kedua kelas tersebut [30].



Gambar 2. 2 Hyperplane yang memisahkan dua kelas positif(+1) dan negatif(-1)

Hyperplane yang ditemukan SVM pada ilustrasi Gambar 2.2 dimana posisi berada diantara dua kelas yang berarti jarak objek data dengan *hyperplane* berbeda dengan kelas berdekatan (terluar) yang diberi tanda bulat kosong dan positif. Objek data terluar dalam SVM itu disebut dengan *support vector*. Untuk menemukan *hyperlplane* yang paling optimal hanya *support vector* inilah yang akan menjadi perhitungan karena sifatnya kritis.

Dalam algoritma Support Vector Machine, terdapat beberapa jenis kernel yang sering digunakan[14], yaitu:

- Kernel Linear, biasanya dataset yang cocok menggunakan kernel ini adalah dataset yang linear.
- Kernel Polynomial, digunakan untuk dataset normal.

- Kernel Radial Basis Function (RBF) atau Gaussian, menjadi kernel yang paling banyak digunakan karena tingginya nilai akurasinya. Biasanya digunakan untuk dataset yang tidak terpisah secara linear.
- Kernel Sigmoid, pengembangan dari jaringan saraf tiruan.

Dalam melakukan klasifikasi terdapat dua proses yaitu:

1. Data *Training*

   Proses ini menggunakan data *train* yang diketahui labelnya untuk dapat dibangun model.

2. Data *Testing*

   Proses ini merupakan mengetahui apakah model yang dibuat sudah akurat ketika proses *training*. Proses ini digunakan untuk dapat melakukan prediksi label-label yang ada.

2.2.9 *Confusion Matrix*

Confusion matrix adalah tabel yang digunakan untuk menganalisis seberapa baik keakuratan suatu metode klasifikasi untuk memprediksi kelas suatu data. *Confusion matrix* divisualisasikan menggunakan tabel dengan melakukan klasifikasi dari jumlah data uji benar dan data uji yang salah[31]. Pada penelitian ini pemilihan metrik untuk mengevaluasi kinerja model yang digunakan pada klasifikasi adalah dengan metrik akurasi, *recall, precision,* dan F1- *Score*. Dari model yang dibuat dan metrik yang dipilih pada data uji kemudian dilakukan evaluasi performa kinerja model yang digunakan.

Ada empat ukuran berbeda dalam metrik yang menentukan kinerja hasilnya[32]:

True Positive (TP) : Ulasan positif yang diklasifikasikan sebagai positif dengan benar(1)

True Negative (TN) : Ulasan negatif yang diklasifikasikan sebagai negatif yang benar(0)

False Positive (FP) : Ulasan positif yang diklasifikasikan sebagai negatif secara salah. Prediksi kelas positif (1) seharusnya kelas negatif(0) (kesalahan Tipe I).
False Negative (FN) : Ulasan negatif yang diklasifikasikan sebagai positif secara salah. Prediksi kelas negatif (0) seharusnya kelas positif(1) (kesalahan Tipe II).

Tabel 2.2 ***Confusion matrix 2x2***

Confussion Matrix		Prediksi	
		Positif "+"	Negatif "-"
Aktual	Positif "+"	TP (True Positive)	FN (False Negative)
	Negatif "-"	FP (False Positive)	TN (True Negative)

Dalam konteks table evaluasi, "0" dan "1" merujuk pada label kelas atau prediksi kelas. Dengan menggunakan istilah "0" dan "1" dalam table evaluasi dapat melihat sejauh mana model dapat membedakan antara kelas positif dan negative dalam suatu masalah klasifikasi. Tabel evaluasi digunakan untuk mengevaluasi kinerja model klasifikasi, dan kelas-kelas ini memiliki makna berikut:

1. 0: label kelas negatif atau kelas yang biasanya digunakan untuk menggambarkan hasil yang tidak diinginkan, salah, atau "tidak ada."
2. 1: lebel kelas positif atau kelas yang biasanya digunakan untuk menggambarkan hasil yang diinginkan, benar, atau "ada."

Confussion matrix selain menunjukan implementasi akhir dari model klasifikasi. Pengukuran confussion *matrix* juga akan menghasilkan suatu nilai untuk akurasi, *precision, recall,* dan *F1-score* yang didefinisikan ke dalam rata-rata presisi dan *recall* tertimbang. Tabel *confussion matrix* juga terdapat metrik evaluasi yang digunakan selain *accuracy, precision, recall* dan *f1-score*. Perbedaan *accuracy, weight average* dan *macro average* terdapat pada cara

menggabungkan metrik -metrik individu untuk mendapatkan metrik yang lebih umum untuk mewakili kinerja model secara keseluruhan[33]. Perbedaan secara detail dapat dilihat sebagai berikut:

1. *Accuracy* (Akurasi)
    - Metrik evaluasi yang paling umum dan sederhana
    - Mengukur sejauh mana model mampu memprediksi dengan benar kelas data yang diamati
    - Dinyatakan dalam persentase dan dihitung sebagai (Jumlah prediksi benar)/(Total sampel).
    - Tidak memperhatikan ketidakseimbangan kelas, yang bisa menjadi masalah jika kelas memiliki distribusi yang tidak seimbang.
2. *Weight Average* (Rata-rata Terimbang)
    - Menghitung metrik individu (seperti precision, recall, atau F1-score) untuk setiap kelas terlebih dahulu.
    - Kemudian, merata-ratakan hasil dengen memperhatikan bobot kelas. Bobot kelas adalah proporsi dari masing-masing kelas dalam data.
    - Ini baik untuk mengatasi masalah ketidakseimbangan kelas karena memberikan bobot lebih besar pada kelas-kelas yang lebih besar.
3. *Macro Average* (Rata-rata Makro)
    - Menghitung metrik evaluasi individu untuk setiap kelas terlebih dahulu
    - Merata-ratakan hasil secara seimbang, tanpa memperhatikan seberapa banyak sampel masing-masing kelas ada dalam data.
    - Memberikan perhatian yang sama pada semua kelas, tidak peduli seberapa besar atau kecil jumlahnya.

Akurasi memberikan gambaran keakuratan model dalam klasifikasi[34]. Perhitungan akurasi dilakukan dengan menjumlahkan data sentimen yang benar dengan dengan total data dan data uji seperti berikut:

$$Accuracy = \frac{(\mathrm{T}P+TN)}{(TP+TN+FN+TN)} \mathrm{x}\ 100\% \qquad (2.4)$$

Presisi mengambarkan akurasi antara data yang diminta dengan hasil prediksi yang diberikan oleh model[34]. Perhitungan presisi dapat dilakukan seperti berikut:

$$Precision\ (P)\ = \frac{\mathrm{TP}}{\mathrm{TP+FP}} \qquad (2.5)$$

Recall atau sensitivity merupakan gambaran keberhasilan model dalam menemukan kembali sebuah informasi[34]. Perhitungan recall dapat dilakukan seperti berikut:

$$Recall\ = \frac{\mathrm{TP}}{\mathrm{TP+FN}} \qquad (2.6)$$

F1-*Score* merupakan gambaran perbandingan rata-rata *precision* dan *recall* yang dilakukan pembobotan. *Accuracy* dapat digunakan sebagai acuan performa algoritma apabila dalam dataset memiliki jumlah data *False Negative* dan *False Positive* yang sangat mendekati. Tetapi apabila tidak mendekati jumlahnya, maka sebaiknya menggunakan F1- *Score* sebagai acuan[34]. Perhitungan F1-*Score* dapat dilakukan sebagai berikut:

$$F1\ -\ Score\ = 2 \times \frac{(\mathrm{Recall} \times \mathrm{Precision})}{\mathrm{Recall} + \mathrm{Precision}} \qquad (2.7)$$