BAB II
TINJAUAN PUSTAKA

2.1. Tinjauan Pustaka

Tinjauan pustaka meliputi penelitian-penelitian terdahulu yang mengarah pada topik yang relevan. Peneliti melakukan kajian terhadap beberapa jurnal yang mengangkat topik analisis sentimen dengan metode yang berbeda-beda diantaranya metode *Naive Bayes*, *Support Vector Machine*, *Random Forest* dan metode lainnya.

Penelitian pada tahun 2020 memprediksi hasil pemilihan presiden Indonesia tahun 2019 menggunakan analisis sentimen dengan metode *SVM* dan *KNN*. Data *tweet* tentang pemilihan presiden Indonesia diambil secara acak pada beberapa tanggal tertentu. Rata-rata akurasi algoritma *SVM* adalah 69,27%, dengan akurasi tertinggi mencapai 76,5%. Sementara itu, nilai rata-rata algoritma KNN adalah 61,3%, dengan akurasi tertinggi mencapai 68,3%. Waktu pelatihan tercepat diperoleh oleh algoritma *KNN*, sedangkan algoritma *SVM* memiliki waktu pengujian tercepat. Hasil prediksi presiden berdasarkan sentimen positif menunjukkan calon nomor urut 01 memperoleh persentase sebesar 67,98%, dan nomor urut 02 memperoleh persentase sebesar 67,79% [15].

Pada tahun 2020, penelitian oleh Prastyo dan tim mengkaji analisis sentimen bahasa Indonesia dengan membandingkan empat fungsi kernel pada *SVM* dengan *TF-IDF*. Fungsi kernel yang paling optimal adalah *RBF* pada *SVM+TF-IDF* dengan fitur 2000, yang mencapai performa terbaik dalam akurasi, presisi, *recall*, dan *f-measure*, masing-masing sebesar 96,61%, 96,70%, 96,58%, dan 96,60%. Fungsi kernel *RBF* menunjukkan kinerja tinggi ketika menggunakan dimensi fitur yang rendah, sementara fungsi kernel *linear* mencapai kinerja terbaik saat menggunakan dimensi fitur yang tinggi [16].

Penelitian tahun 2022 mengenai analisis sentimen terhadap aplikasi Peduli Lindungi di *Google Play* menggunakan metode *Random Forest* dengan *SMOTE*.

Hasilnya menunjukkan bahwa implementasi *Random Forest* dengan *SMOTE* menghasilkan akurasi 71%, recall 70%, dan presisi 70%. Sementara itu, implementasi *Random Forest* tanpa *SMOTE* menghasilkan akurasi 60%, *recall* 57%, dan presisi 55%. Penggunaan *SMOTE* meningkatkan akurasi sebesar 11%, *recall* sebesar 13%, dan presisi sebesar 15% [17].

Pada tahun 2018, dilakukan penelitian menggunakan pendekatan *hybrid random forest* dan *support vector machin*e *(RF-SVM)* untuk mengklasifikasikan ulasan produk di *marketplace Amazon*. Hasilnya menunjukkan bahwa metode *Hybrid RF-SVM* menghasilkan nilai *precision*, *recall*, dan *F-measure* masing-masing 83,4%; 84,4%; dan 83,4%, menunjukkan stabilitas yang baik [10].

Penelitian pada tahun 2017 meneliti tentang analisis jaringan semantik sentimen vaksin di media sosial online dengan dataset diambil dari data artikel vaksin di *Twitter* Amerika Serikat. Hasilnya menunjukkan rata-rata dari sentralitas derajat, sentralitas antara, sentralitas kedekatan, dan sentralitas vektor *eigen* pada sentimen vaksin positif, negatif, dan netral [18].

Pada tahun 2021, dilakukan penelitian tentang analisis sentimen *Twitter* terhadap vaksin Covid-19 di Filipina dengan menggunakan metode *naive Bayes.* Hasilnya menunjukkan mayoritas *tweet* di Filipina (83%) adalah positif terhadap vaksin Covid-19, 9% netral, dan 8% negatif. Akurasi *naive Bayes* melalui operator *RapidMiner* mencapai 81,77%, melebihi akurasi studi sebelumnya [8].

Penelitian pada tahun 2019 menggabungkan metode klasifikasi *naive Bayes* dengan *k-means* untuk menganalisis sentimen ulasan produk di e-commerce *Shopee*. Hasilnya menunjukkan bahwa kombinasi pengklasifikasi *k-means* dan *naive Bayes* tanpa manual memiliki akurasi 77,12%, sedangkan proses yang dilakukan secara manual memiliki akurasi 56,86% [9].

Penelitian pada tahun 2020 menganalisis sentimen pada *review* produk tokopedia dengan metode *random forest*. Metode ini menghasilkan akurasi sebesar 97,38% yang menunjukkan bahwa metode *random forest* dapat memprediksi ulasan produk Tokopedia dengan baik [19].

Pada tahun 2018, Yuling Chen dan Zhi Zhang membangun model analisis sentimen dengan menggabungkan keunggulan metode *CNN* dan *SVM*. Hasilnya

menunjukkan bahwa model *CNN-SVM* memiliki akurasi yang jauh lebih tinggi daripada model terbaik lainnya dalam klasifikasi teks sentimen [20].

Penelitian tahun 2021 menganalisis sentimen bahasa Urdu dengan menggunakan metode *Deep Learning*. Metode ini mencapai skor *F1* tertinggi 82,05% dengan menggunakan kombinasi fitur kata *n-gram* dengan *logistic regression* [21].

Tabel 2. 1 *Literature review* dari penelitian sebelumnya.

No.	Penulis	***Comparing***	***Contrasting***	***Criticize***	***Synthesize***	***Summarize***
1.	*Comparing Sentiment Analysis of Indonesian Presidential Election 2019 with Support Vector Machine and K-Nearest Neighbor Algorithm* (Fiki Firmansyah, Wildan Budiawan Zulfikar, Dian Sa'adillah Maylawati, Nunik Destria Arianti, Lia Muliawaty, Muhammad Andi Septiadi, Muhhammad Ali Ramdhani, 2020)	Memprediksi hasil pemilihan presiden Indonesia 2019 berdasarkan analisis sentimen di *twitter* dengan menggunakan metode *SVM* dan *KNN*	Penggunaan algoritma *SVM* dan algoritma *KNN*, Pembobotan menggunakan TF-IDF.	Waktu pelatihan Algoritma *SVM* lebih lama karena proses penentuan *support vector* dan *hyperplane* untuk mendapatkan model. Sedangkan pelatihan *KNN* hanya menyimpan data yang telah diberi label sebelumnya sehingga membutuhkan waktu yang lebih sedikit untuk melakukan pelatihan.	Data *tweet* acak pada tanggal 27 Februari 2019, 28 Februari 2019, 7 April 2019, 9 April 2019, 11 April 2019, dan 12 April 201. Data dibagi menjadi data latih dan uji, kemudian data tersebut di-*preprocessing* dan dilakukan pembobotan dengan TF-IDF selanjutnya dilakukan prediksi dan evaluasi dari metode *SVM* dan *KNN*.	Rata-rata akurasi algoritma *SVM* adalah 69,27 dengan akurasi tertinggi 76,5%, sedangkan nilai rata-rata algoritma *KNN* adalah 61,3% dengan akurasi tertinggi 68,3%. Waktu pelatihan tercepat didapatkan oleh algoritma *KNN*, sedangkan algoritma *SVM* mendapatkan waktu pengujian tercepat. Hasil prediksi presiden berdasarkan sentimen positif yaitu calon nomor urut 01 memperoleh persentase sebesar 67,98% dan nomor urut 02 memperoleh persentase sebesar 67,79%.
2.	*Indonesian Sentiment Analysis: An Experimental Study of Four Kernel Functions on SVM Algorithm with TF-IDF(*Pulung Hendro Prastyo, Igi Ardiyant, Risanuri Hidayat,	Membandingkan empat fungsi kernel pada *SVM* dengan *TF-IDF*, seperti kernel *Polynomial*, *Sigmoid*, *Linear*, dan Radial Basis Function (*RBF*). *TF-IDF* digunakan sebagai ekstraksi fitur dan	Penggunaan algoritma *SVM* dengan kernel yang berbeda-beda dengan pembobotan *TF-IDF*.	Masih adanya perbedaan hasil yang bervariasi untuk SVM saat digunakan untuk analisis sentimen pada data Twitter terkait isu undang-undang omnibus di Indonesia	Menggunakan perpustakaan *GetOldTweets3* diperoleh data crawling sebanyak 14.398 *tweet*. Kata kunci yang digunakan adalah #omnibuslaw dan omnibus law. Kemudian,	Berdasarkan hasil percobaan, fungsi kernel *RBF* pada *SVM+TF-IDF* menggunakan fitur 2000 mendapat performa terbaik dari semua pengujian fitur dalam akurasi, presisi, *recall*, dan *f-measure* dengan masing-masing nilai

No.	Penulis	*Comparing*	*Contrasting*	*Criticize*	*Synthesize*	*Summarize*
	2020*)*	seleksi untuk meningkatkan kinerja *SVM*.			data yang tidak diinginkan dan data duplikat dihapus. Menggunakan 4.000 *tweet* berlabel, dimana 2.000 *tweet* adalah sentimen positif, dan 2.000 *tweet* adalah sentimen negatif. Selanjutnya dilakukan preprocessing data, pembobotan dengan tf-idf, kemudian dilakukan pengembangan dan perbandingan 4 kernel *SVM*, dan melakukan evaluasi.	96.61%, 96.70%, 96.58%, 96.60%. Fungsi kernel *RBF* memperoleh kinerja tinggi saat menggunakan dimensi fitur yang rendah. Sementara itu, fungsi kernel *Linear* mencapai kinerja terbaik saat menggunakan dimensi fitur yang tinggi.
3.	*Sentiment Analysis of the PeduliLindungi on Google Play using the Random Forest Algorithm with SMOTE* (Muhammad Rizky Pribadi, Danny Manongga, Hindriyanto Dwi Purnomo, Hendry, Iwan Setyawan, 2022)	Membandingkan dua eksperimen klasifikasi sentimen terhadap aplikasi peduli lindungi dengan metode *Random Forest* tanpa *SMOTE* dan *Random Forest* dengan *SMOTE*	Penggunaan metode *Random Forest* dengan *SMOTE* dan tanpa *SMOTE*	Rasio antara sentimen negatif dan positif adalah 60% dengan sentimen positif yang sedikit. Proses klasifikasi dilakukan terlalu cepat dapat menyebabkan hasil klasifikasi yang kurang baik. Oleh karena itu, untuk mengatasi masalah tersebut digunakan metode *SMOTE*. Setelah menerapkan metode *SMOTE*,	Dataset dikumpulkan dari komentar pengguna aplikasi Peduli Lindungi di Google Play mulai dari 15 Februari sampai 15 Maret 2022. Dataset di-*preprocessing* dan dilakukan pembobotan dengan *TF-IDF*, selanjutnya dilakukan optimasi dengan *SMOTE* untuk mengatasi ketidakseimbangan dataset. Kemudian dilakukan klasifikasi dengan *Random Forest*	Implementasi *Random Forest* dan *SMOTE* menghasilkan akurasi sebesar 71%, *recall* 70%, dan presisi 70%. Sedangkan implementasi *Random Forest* tanpa *SMOTE* menghasilkan akurasi 60%, recall 57%, dan presisi 55%. Oleh karena itu, penerapan *SMOTE* dapat meningkatkan akurasi sebesar 11%, *recall* sebesar 13%, dan presisi sebesar 15%.

No.	Penulis	*Comparing*	*Contrasting*	*Criticize*	*Synthesize*	*Summarize*
				pembagian tiap kelas menjadi 33%.	dan evaluasi pada klasifikasi dengan *Random Forest* dengan *SMOTE* dan tanpa *SMOTE*.	
4.	*Random Forest and Support Vector Machine based Hybrid Approach to Sentiment Analysis* (Yassine AL AMRANI, Mohamed LAZAAR, Kamal Eddine EL KADIRI, 2018)	Menggunakan pendekatan hybrid *RF* dan *SVM* untuk mengklasifikasikan ulasan produk amazon	Penggunaan pendekatan hybrid *RF* dan *SVM*	Masih terdapat beberapa kekurangan dalam klasifikasi menggunakan metode *Random Forest (RF)* dan *Support Vector Machine (SVM.*	Data latih dan uji diambil dari dataset produk amazon yang terdiri dari 500 positif dan 500 negatif. Pelatihan dan pengujian dengan metode *Cross Validation* dengan nilai fold adalah 10. Data dilakukan *feature extraction* dan dilakukan pengklasifikasian dengan metode *RF*, *SVM*, dan *RFSVM*	Dengan menggunakan pendekatan hybrid *RFSVM* dalam kasus klasifikasi ulasan produk di amazon mendapatkan hasil yang lebih baik karena memanfaatkan keunggulan masing-masing metode klasifikasi RF tradisional dan juga *SVM*. Dengan *precision*, *recall*, dan *F-measure* masing-masing 83,4%; 84,4%, dan 83,4% dengan stabilitas yang baik.
5.	*Semantic Network Analysis of Vaccine Sentiment in Online Social Media* (Gloria J. Kang, Sinclair R. Ewing-Nelson, Lauren Mackey, James T. Schlitt, Achla Marathe, Kaja M. Abbas, Samarth Swarup, 2017)	Membangun jaringan semantik informasi vaksin dengan menggunakan *ChatterGrabber* untuk mengambil sampel *tweet*, kemudian dianalisis dan divisualisasikan.	Analisis topologi sentimen, membandingkan perbedaan semantic, dan mengidentifikasi konsep yang paling menonjol dalam jaringan yang mengekspresikan sentimen vaksin positif, negatif, dan netral.	Penentuan sentimen netral sulit dilakukan karena dokumen menampilkan campuran sikap positif dan negatif, dan tidak benar-benar netral vaksin. Karena perilaku kesehatan didasarkan pada berbagai keyakinan dan sikap yang berubah dari waktu ke waktu, kategori sentimen vaksin sulit	26.389 tweet dikumpulkan antara 16 April 2015 dan 29 Mei 2015 diperoleh 8.416 tautan web unik. Untuk menggeneralisasi temuan dari kumpulan artikel vaksin populer yang representatif, dilakukan penyaringan 100 tautan yang paling banyak dibagikan lalu diambil 50 sampel secara acak untuk dianalisis. Kemudian dilakukan	50 data yang telah terpilih dari 26.389 data *tweet* yang dikumpulkan dengan *Chatter Grabber*. jaringan sentimen vaksin positif, negatif, dan netral didapatkan rata-rata dari sentralitas derajat, sentralitas antara, sentralitas kedekatan, dan sentralitas vektor eigen secara berturut-turut: positif: 0,0061; 0,006; 0,2292; 0,0626 negatif: 0,0027; 0,0033;

No.	Penulis	*Comparing*	*Contrasting*	*Criticize*	*Synthesize*	*Summarize*
				digambarkan karena tidak ada sebagai kelompok yang terpolarisasi.	pengkodean secara manual yang memiliki sentiment positif, negative, dan netral terhadap vaksin. Selanjunya dilakukan pembangunan jaringan sentiment vaksin, visualisasi dinamis dan analisis.	0,2161; 0,0318 netral: 0,0149; 0,0342; 0,1533; 0,0975.
6.	*Twitter sentiment analysis towards covid-19 vaccines in the Philippines using naïve bayes* (Charlyn Villavicencio, Julio Jerison Macrohon, X. Alphonse Inbaraj, Jyh-Horng Jeng, Jer-Guang Hsieh, 2021)	Menggunakan model *Naïve Bayes* untuk mengklasifikasikan *tweet* berbahasa Inggris dan Filipina ke dalam polaritas positif, netral, dan negatif melalui perangkat lunak ilmu data *RapidMiner*.	Penggunaan algoritma *naïve bayes*, pembobotan dengan *TF-IDF*	Keterbatasan RapidMiner gratis dalam hal jumlah data yang dapat diimpor, model *Extract Sentiment* di RM tidak menyertakan bahasa Filipina, jadi para peneliti harus membubuhi keterangan tweet yang dikumpulkan dengan tangan.	11.974 tweet dikumpulkan dari 1-31 Maret 2021, kemudian disaring menjadi 993 tweet. Selanjutnya data ini dilabeli dan dilakukan *preprocessing*, pembobotan dengan *TF-IDF*, klasifikasi dengan *naïve bayes*. Kemudian hasil klasifikasi di evaluasi.	Menghasilkan 81,77% yang melebihi akurasi studi analisis sentimen pada periode waktu berdekatan yang menggunakan data Twitter di Filipina.
7.	*Sentiment Analysis of Product Reviews as A Customer Recommendation Using the Naive Bayes Classifier Algorithm* (Taqwa Hariguna, Wiga Maulana Baihaqi, Aulia Nurwanti, 2019)	Analisis Sentimen Ulasan Produk sebagai Rekomendasi Pelanggan menggunakan Algoritma Klasifikasi *Naive Bayes*	Penggunaan algoritma *Naïve Bayes* dan dikombinasikan dengan *K-means*	Proses pengelompokan sentimen menggunakan K-means belum sepenuhnya maksimal karena masih ada terdapat kesalahan dalam memasukkan kalimat ke dalam	Data diperoleh dari komentar positif dan review negatif terhadap produk berbahasa Indonesia di website e-commerce sebanyak 153 data komentar. Data tersebut kemudian dilakukan *preprocessing* Setelah itu dilakukan	Ketepatan hasil yang diperoleh dari kombinasi pengklasifikasi K-means dan naive Bayes tanpa manual dalam menganalisis sentimen ulasan produk shopee adalah sebesar 77,12% sedangkan proses yang dilakukan dengan manual mendapatkan hasil

No.	Penulis	*Comparing*	*Contrasting*	*Criticize*	*Synthesize*	*Summarize*
				kelas positif dan kelas negatif. Untuk itu perlu adanya bantuan manual untuk mengoreksi pengelompokan kelas positif dan kelas negatif.	proses clustering dengan K-means antara lain cluster 0 atau negatif berjumlah 116 data komentar review produk, dan cluster 1 atau positif berjumlah 37 data komentar review produk, kemudian dilakukan koreksi secara manual menjadi 89 negatif dan 64 positif. Klasifikasi dan evaluasi dengan naïve Bayes.	akurasi sebesar 56,86%.
8.	*Sentiment Analysis on Tokopedia Product Online Reviews Using Random Forest Method* (Stephenie, Budi Warsito, Alan Prahutama, 2020)	Menggunakan metode *Random Forest* dan 10 lipat cross-validation. Pelabelan data menggunakan Lexicon. Visualisasi hasil pelabelan kemudian dilakukan dengan menggunakan grafik batang dan *word cloud* pada masing-masing kelas sentimen untuk mencari informasi yang dianggap penting dan paling banyak dibicarakan	Penggunaan metode Random Forest dengan parameter mtry = 73 dan ntree = 50 untuk klasifikasi sentimen.	Dalam proses pelabelan ada banyak review deskripsi konten dan rating yang tidak searah oleh karena itu perlu dilakukan pelabelan data berdasarkan isi ulasan.	Dataset yang digunakan adalah review produk Tokopedia berbahasa Indonesia dengan jumlah total 40.607 review. Dilakukan *preprocessing* data membuat jumlah ulasan menurun menjadi 23.639, data ini kemudian dilabeli menjadi sentimen positif, negatif, dan netral. Selanjutnya dilakukan pembobotan dengan TF-IDF, visualisasi ulasan dengan menggunakan diagram frekuensi kata terbanyak dan *word cloud.* Selanjutnya dilakukan klasifikasi	Hasil pengujian menunjukkan bahwa akurasi Metode Random Forest dengan parameter mtry = 73 dan ntree = 50 adalah 97,38% yang mengarah pada kesimpulan bahwa Metode *Random Forest* dapat memprediksi ulasan produk Tokopedia dengan baik dimana Semakin besar akurasi, semakin baik kinerja model klasifikasi.

No.	Penulis	*Comparing*	*Contrasting*	*Criticize*	*Synthesize*	*Summarize*
					dengan *Random Forest* dan evaluasi model.	
9.	*Research on text sentiment analysis based on CNNs and SVM* (Yuling Chen dan Zhi Zhang, 2018)	Menggunakan metode gabungan *CNN-SVM* untuk klasifikasi sentimen emosi pada dataset NLPCC2014.	Penggunaan *CNN* sebagai automatic feature learner dan *SVM* sebagai pengklasifikasi teks sentimen emosi.	*CNN* dapat mengekstrak representasi fitur yang bermakna dari sampel input secara efektif namun memiliki klasifikasi yang lemah pada data yang dipisahkan secara nonlinier. Untuk itu dibutuhkan model pembelajaran mesin untuk melakukan klasifikasi pada data tersebut, salah satu metode yang baik untuk itu adalah *SVM*.	Dataset yang digunakan adalah *NLPCC2014 emotional analysis* yang dievaluasi dengan *deep learning*. Terdiri dari 10.000 data latih berupa 5.000 data polaritas emosional positif dan 5.000 data polaritas emosional negatif. 2.500 data uji berupa 1.250 data polaritas emosional positif dan 1.250 negatif. Data-data ini kemudian di proses, segmentasi kata di filter supaya bersih dan data terfilter kemudian di dilatih oleh *Word2vec* menjadi vektor fitur emosi 300 dimensi. Selanjutnya dilakukan klasifikasi dan evaluasi model.	Hasil penelitian menunjukkan bahwa kombinasi antara *CNN* dan *SVM* dalam tugas klasifikasi sentimen emosi lebih baik dari pada metode NLPCC_SCDL_Best dan metode *CNN* saja. Hasil evaluasi penggabungan *CNN* dan *SVM* memiliki nilai Presisi, *Recall*, dan *F-Measure* secara berturut-turut adalah 0,890; 0,886; 0,888 untuk sentimen positif dan 0,886; 0,891; 0,889 untuk sentimen negatif.
10.	*Urdu Sentiment Analysis with Deep Learning Methods* (Lal Khan, Ammar Amjad, Noman Ashraf, Hsien-Tsung Chang, Alexander	Analisis Sentimen dalam Bahasa Urdu dengan Metode *Deep Learning*	Penggunaan model pembelajaran mesin RF, NB, *SVM*, AdaBoost, MLP, LR dan *deep learning 1D-CNN*, *LSTM*.	Kekurangan sumber daya linguistik dan bahasa seperti leksikon dan kumpulan menyulitkan penerapan metode	Dataset yang digunakan dikumpulkan dari berbagai sumber ulasan tentang politik, film, drama urdu, acara bintang-bintang TV, dan olahraga dikumpulkan	Hasil penelitian menunjukkan bahwa kombinasi fitur kata n -gram dengan *LR* mengungguli metode pengklasifikasian lainnya untuk menganalisis sentimen memperoleh skor

No.	Penulis	*Comparing*	*Contrasting*	*Criticize*	*Synthesize*	*Summarize*
	Gelbukh, 2021)			analisis sentimen seperti ketersediaan leksikon dan kumpulan data. Untuk mengurangi kekurangan ini, penelitian ini menekankan pada pembuatan dataset bahasa Urdu yang berisi kalimat-kalimat yang dipasang pada enam domain berbeda.	oleh ahli bahasa urdu selama 3 bulan. Dengan review negatif sebanyak 4.758, positif sebanyak 4.843. kemudian dilakukan *preprocessing, feature selection, word embedding* dengan *fast Text*, setelah itu dilakukan klasifikasi dengan *SVM, NB, RF, AdaBoost, MLP, 1D-CNN,* dan *LSTM* untuk menemukan keefektifan korpus. Selanjutnya dilakukan evaluasi keefektifan model menggunakan *Recall, Precision* dan *F1-Measure*	F1 tertinggi 82,05% dengan menggunakan kombinasi fitur kemudian disusul oleh *SVM* sebesar 81,47% dan memiliki kinerja rata-rata yang paling baik dari semua pengklasifikasi lainnya.

Berdasarkan Tabel 2.1 *Literature review* dan dari penelitian sebelumnya, penggunaan algoritma *SVM* dalam melakukan klasifikasi teks mempunyai akurasi yang baik, jurnal oleh Fiki Firmansyah et al mengenai analisis sentimen pemilihan presiden Indonesia tahun 2019 menjadi referensi penggunaan algoritma *support vector machine* (*SVM*), jurnal oleh Pulung Hendro et al tentang analisis sentimen menggunakan *SVM* dengan 4 kernel sebagai referensi penggunaan kernel *SVM*, jurnal oleh Muhammad Rizky Pribadi et al yang membandingkan hasil klasifikasi sentimen dengan menggunakan *SMOTE* dan tanpa *SMOTE* sebagai referensi penyeimbangan dataset, dan jurnal lainnya sebagai referensi dalam melakukan analisis sentimen.

2.2. Dasar Teori

Pada bab ini, akan dibahas teori-teori yang terkait dengan topik penelitian. Teori ini termasuk Analisis Sentimen, *Support Vector Machine,* parameter evaluasi, *Twitter*, *Youtube*, dan teori lainnya yang relevan.

2.2.1 Analisis Sentimen

Analisis Sentimen *(Sentiment Analysis)* merupakan suatu kegiatan untuk mengetahui dan mengklasifikasikan emosi seseorang seperti emosi baik, buruk, ataupun emosi netral yang dimuat dalam suatu tulisan atau komentar seseorang [22]. Analisis sentimen ini sering dipakai untuk mengetahui persentase dari suatu sentimen yang positif, negatif, maupun yang netral pada suatu *review* produk, komentar-komentar yang ada di sosial media, maupun komentar-komentar yang ada di suatu berita atau artikel [19].

2.2.2 *Crawling*

Merupakan proses pengambilan data dalam jumlah banyak yang ada pada suatu platform kemudian diunduh ke dalam file lokal komputer. *Crawling* data ini dilakukan untuk ekstraksi data yang ada pada *website*, media sosial, dokumen, ataupun file berupa *tweet*, ulasan, komentar, spesifikasi produk dan lain lain. Data-data hasil crawling ini nantinya akan diproses sesuai dengan kebutuhan dari setiap kasus. Pada media twitter dapat dilakukan crawling data yaitu mengumpulkan *tweet-tweet* dalam

jumlah yang banyak dengan menggunakan *API* dari *Twitter* [23], selain dari *Twitter crawling* data juga dapat dilakukan pada *Youtube* berupa komentar dari pengguna *youtube* dengan menggunakan *tools* tertentu [24].

2.2.3 *Preprocessing*

Preprocessing data merupakan tahap awal yang dilakukan untuk menyempurnakan data yang telah di *crawling* agar lebih mudah diproses. Beberapa langkah yang dilakukan dalam preprocessing data meliputi *case dolding*, yaitu mengubah semua huruf menjadi huruf kecil; *tokenization*, yaitu memecah dokumen menjadi bagian-bagian kata yang disebut token; *filtering*, yaitu menghapus tanda baca dan karakter non-alfabet; *stopword removal*, yaitu mengambil kata-kata yang dianggap penting atau menyingkirkan kata-kata yang dianggap tidak terlalu memiliki arti penting dalam proses *text mining*; *stemming* yaitu mentransformasikan kata menjadi kata dasar dengan menghapus imbuhan, ataupun lemmatisasi yaitu mentransformasikan kata menjadi bentuk dasar berdasarkan konteks dan makna dimaksudkan[25], [16].

2.2.4 *Wordcloud*

Wordcloud adalah gambar yang menampilkan daftar kata-kata yang terdapat dalam sebuah teks. Semakin sering sebuah kata muncul dalam teks, maka akan semakin besar ukuran gambar kata tersebut. Sebaliknya, semakin jarang sebuah kata muncul dalam teks, maka akan semakin kecil ukuran gambar kata tersebut [19].

2.2.5 Pelabelan

Pelabelan merupakan proses pemberian label, kategori, atau tanda tertentu pada data atau teks untuk tujuan analisis, klasifikasi, atau pemahaman lebih lanjut. Dalam konteks analisis data atau text mining, pelabelan dapat dilakukan secara manual oleh manusia atau menggunakan teknik pembelajaran mesin, *deep learning* untuk memberikan klasifikasi otomatis berdasarkan data pelatihan yang telah diberi label sebelumnya [26].

2.2.6 *Indonesian RoBERTa Sentiment Classifier Inference*

Indonesian RoBERTa Sentiment Classifier Inference adalah model yang difokuskan pada analisis sentimen dalam teks berbahasa Indonesia. Model ini berbasis pada arsitektur *transformer RoBERTa* yang telah menjalani pelatihan luas dengan menggunakan dataset teks bahasa Indonesia yang sangat besar. Proses pelatihan model melibatkan tahap pre-training, di mana *RoBERTa* dilatih secara tidak terawasi pada dataset teks besar untuk memahami pola-pola umum bahasa, dan tahap *fine-tuning*, di mana model dilatih pada dataset yang telah ditandai dengan sentimen positif, negatif, atau netral. Selama fase inferensi, teks berbahasa Indonesia yang diberikan dipecah menjadi token-token, dimasukkan ke dalam arsitektur *RoBERTa*, dan menghasilkan representasi vektor teks.

Representasi vektor ini selanjutnya diarahkan ke lapisan klasifikasi yang mengeluarkan prediksi sentimen untuk teks tersebut, yakni sentimen positif, negatif, atau netral. Dengan dukungan dari dataset pelatihan yang besar dan beragam, *Indonesian RoBERTa Sentiment Classifier* mampu memahami kompleksitas bahasa Indonesia dalam konteks analisis sentimen, dan memberikan hasil analisis sentimen yang akurat, sangat berperan dalam berbagai aplikasi, termasuk pemantauan citra merek, analisis umpan balik pelanggan, dan deteksi sentimen dalam konten media sosial berbahasa Indonesia [26].

2.2.7 Pembobotan

Pembobotan adalah suatu cara untuk merubah masukan data menjadi suatu fitur vektor [15]. Setiap file teks mempunyai banyak macam kata untuk menyusun sebuah kalimat-kalimat ataupun paragraf. Untuk itu perlu dilakukan pemberian bobot untuk setiap kata (*term*) pada suatu dokumen yang tentunya memiliki tingkat kepentingan atau kontribusi yang berbeda-beda agar dapat ditentukan kategori atau kelas kata tersebut [16], [25].

Teknik pembobotan yang biasa dipakai pada pembobotan kata adalah *Term Frequency-Inverse Document Frequency (TF-IDF)*. *TF-IDF* adalah suatu metode pembobotan yang menggabungkan dua konsep yaitu *Term*

Frequency dan *Document Frequency* [15]. Perhitungan *TF-IDF* dapat dilihat pada persamaan berikut:

$$TF\text{-}IDF_{(t,d)} = TF_{(t,d)} \times IDF_{(t)} \quad (2.1)$$

$$TF_{(t,d)} = \frac{f_{(t,d)}}{n_{(d)}} \quad (2.2)$$

$$IDF_{(t)} = \log(\frac{N}{DF_t}) \quad (2.3)$$

Dengan:

t	=	kata atau *term* yang ingin dihitung nilai *TF-IDF*-nya dalam dokumen
d	=	dokumen di dalam *dataset*
$n_{(d)}$	=	total jumlah kata dalam dokumen *d*
N	=	jumlah seluruh dokumen dalam *dataset*
DF_t	=	umlah dokumen yang mengandung kata *t* dalam seluruh koleksi dokumen
$TF_{(t,d)}$	=	*Term Frequency* dari kata dalam dokumen
$IDF_{(t)}$	=	*Inverse Document Frequency* dari kata dalam seluruh koleksi dokumen

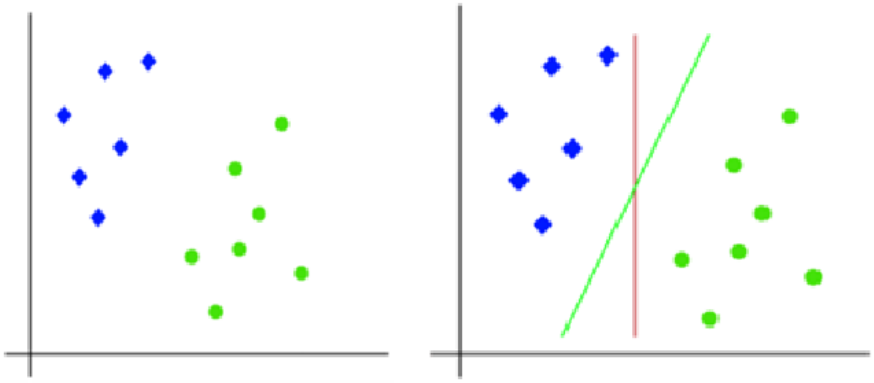
2.2.8 *Support Vector Machine*

Support Vector Machine (SVM) adalah metode pengklasifikasian yang ditemukan pada tahun 1992 oleh Boser, Guyon, dan Vapnik saat *Annual Workshop on Computational Learning Theory*. Metode ini merupakan hasil penggabungan atau kombinasi dari beberapa teori komputasi yang sudah ada sebelumnya. *SVM* berbeda dengan metode jaringan saraf *(neural network),* yang bertujuan menemukan hyperplane pemisah antar kelas.

support vector machine (SVM) adalah teknik pembelajaran mesin yang efektif untuk melakukan klasifikasi dengan kinerja generalisasi yang baik. *SVM* termasuk dalam kelas pembelajaran terarah (*supervised learning)* dan akan menemukan hyperplane yang memisahkan dua kelas pada input space [15]. Dalam *SVM*, ada istilah *support vector* yang merujuk pada dua

data kelas yang berbeda dengan jarak terdekat, *hyperplane* yang merupakan garis pembatas antara kedua *support vector*, dan margin yang merupakan jarak antara *support vector* dan *hyperplane* [20]. Margin yang dibuat harus maksimum untuk mengantisipasi data yang mirip dengan kelas lain.

Data yang *linear*, kita dapat menggunakan *SVM linear* yang dapat memisahkan dataset dengan garis lurus tunggal menjadi dua kelas secara *linear*, seperti yang terlihat pada Gambar 2.1.



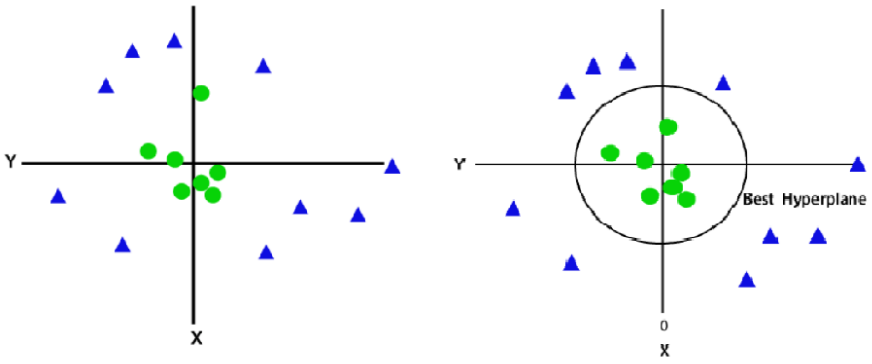
Gambar 2. 1 Ilustrasi *Dataset Linear* [27]

Rumus yang dapat digunakan untuk menghitung *SVM linear* adalah:

$$f(x) = wx + b \tag{2.4}$$

Dengan $f(x)$ adalah fungsi klasifikasi untuk memprediksi kelas data x, w adalah bobot untuk mengukur tingkat kontribusi dalam data x, x adalah vektor fitur dari data yang akan diklasifikasi, dan b adalah bias atau konstanta yang digunakan untuk menyesuaikan fungsi klasifikasi.

Untuk data yang tidak dapat dibagi dengan garis lurus atau tidak *linear*, kita dapat menggunakan *SVM non-linear*. Ilustrasi *SVM non-linear* dapat dilihat pada Gambar 2.2.



Gambar 2. 2 Ilustrasi *Dataset non-Linear* [27]

Rumus yang dapat digunakan untuk menghitung *SVM non-linear* adalah:

$$f(x) = w\varphi(x) + b \tag{2.5}$$

Dengan $f(x)$ adalah fungsi klasifikasi untuk memprediksi kelas data x, w adalah bobot untuk mengukur tingkat kontribusi dalam data x, x adalah vektor fitur dari data yang akan diklasifikasi, φ adalah fungsi kernel untuk mengubah data x dari ruang fitur asli ke ruang fitur dengan dimensi yang lebih tinggi, dan b adalah bias atau konstanta yang digunakan untuk menyesuaikan fungsi klasifikasi.

Support Vector Machine (*SVM*) juga memiliki beberapa kernel yang dapat meningkatkan metode *SVM* seperti kernel *Polynomial*, *Sigmoid*, *Linear*, dan Radial Basis Function (*RBF*) [16].

Tabel 2. 2 Rumus kernel *SVM*

Kernel	Rumus
Polynomial	$\mathrm{k(x,y)} = (\mathrm{x.y} + \mathrm{c})^{\mathrm{d}}$
Sigmoid	$k(x,y) = \tanh(\gamma x.y + c)$
Linear	$k(x,y) = x.y + c$
Radial Basis Function (RBF)	$k(x,y) = \exp\left(-\gamma\left\|\|x-y\right\|\right\|^2\right)$

di mana:

$x\ dan\ y$	:	vektor fitur dari data yang akan diprediksi.
c	:	konstanta yang digunakan untuk menyesuaikan fungsi kernel.
d	:	derajat dari *polynomial*.
γ	:	parameter yang digunakan untuk mengontrol tingkat kompleksitas dari kernel.
$\lvert\lvert x-y\rvert\rvert^2$	:	*euclidean distance* atau jarak antara x dan y.

2.2.9 Parameter Evaluasi

Penentuan performa terbaik dari suatu model dapat dilakukan dengan cara membandingkan evaluasi parameter dari tiap-tiap model. Untuk mengevaluasi performa suatu model klasifikasi, dapat digunakan beberapa metrik seperti akurasi, presisi, sensitivitas, spesifisitas, dan *F-1 score*. Konsep dasar yang digunakan dalam mengukur performa model ini adalah *Confusion Matrix*, yang menunjukkan data yang diprediksi benar atau salah berdasarkan nilai sebenarnya. *True Positive (TP*) adalah data yang diprediksi positif dan benar, True Negative (TN) adalah data yang diprediksi negatif dan benar, *False Positive (FP)* adalah data yang diprediksi positif namun salah, dan *False Negative (FN)* adalah data yang diprediksi negatif namun salah [10], [17], [28]. Pemilihan metrik tergantung pada jenis studi kasus yang sedang dimodelkan.

a. Akurasi

Akurasi dihitung dengan menghitung rasio antara jumlah prediksi yang benar (baik data positif maupun data negatif) terhadap total data [19][28]. Akurasi dapat dihitung dengan persamaan berikut:

$$Akurasi = \frac{TP+TN}{TP+FP+TN+FN} \quad (2.4)$$

Akurasi lebih cocok digunakan jika perbandingan jumlah label suatu data relatif sama, jika perbandinganya tidak sama, maka dapat ditinjau menggunakan metrik lain.

b. Presisi

Presisi merupakan rasio antara jumlah prediksi benar positif terhadap total data dari hasil prediksi positif [10], [28]. Presisi dapat dirumuskan dengan:

$$Presisi = \frac{TP}{TP+FP} \quad (2.5)$$

Presisi sangat cocok jika kita ingin menghindari data yang diprediksi positif namun bernilai salah (*FP*).

c. *Recall* atau Sensitivitas

Recall atau Sensitivitas merupakan rasio antara jumlah prediksi benar positif terhadap total data yang berlabel positif [10], [28]. *Recall* atau Sensitivitas dapat dirumuskan dengan:

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

Recall sangat cocok jika kita ingin menghindari data yang diprediksi negatif namun bernilai salah (*FN*).

d. F-1 *Score*

Dari penjelasan sebelumnya, presisi sangat cocok jika kita ingin menghindari data yang diprediksi positif namun bernilai salah (*FP*) dan *Recall* sangat cocok jika kita ingin menghindari data yang diprediksi negatif namun bernilai salah (*FN*). Hal tersebut dapat menyebabkan ambiguitas, untuk menghindari ambiguitas tersebut maka dapat menggunakan metrik *F-1 Score* yang dapat menghindari nilai *FP* dan *FN* secara bersamaan. Metrik ini diperoleh dari rataan harmonik presisi dan sensitivitas [10], [28]. Rataan harmonik merupakan rataan yang didapat dengan mengubah semua data kedalam bentuk pecahan terlebih dahulu. *F-1 scores* atau *F-measures* dapat dirumuskan dengan:

$$F-1\ Score = 2\ \times \frac{Presisi \times Recall}{Presisi + recall} \quad (2.7)$$

2.2.10 *Twitter*

Twitter merupakan salah satu aplikasi tepatnya media sosial yang dapat digunakan penggunanya untuk melihat dan memberikan suatu pesan berupa teks, gambar, ataupun suatu video yang disebut dengan tweet. *Twitter* ini cukup berbeda dibandingkan dengan media sosial lainnya yang membebaskan penulisnya menulis sebebasnya tanpa batasan jumlah karakter sedangkan di *twitter* kita hanya dapat menuliskan 280 karakter yang berisikan perasaan, kegelisahan, ataupun cuitan kita. semua *tweet* yang kita tuliskan atau kita bagikan dapat dilihat oleh semua pengguna lainnya karena *twitter* bersifat publik, namun kita sebagai pengguna *twitter* juga dapat membatasi orang-orang yang dapat melihat *tweet* kita baik itu hanya bisa dilihat teman saja atau *follower* kita dan bisa juga diatur untuk dilihat semua pengguna lainnya [29].

2.2.11 *Youtube*

Situs web *Youtube* memungkinkan pengguna untuk menonton, mengunggah, dan membagikan video secara publik. Pendiri *Youtube* yaitu Steve Chen, Chad Hurley, dan Jaweb Karim, merupakan mantan karyawan Paypal yang menciptakan situs ini pada bulan Februari 2005. Kini, *Youtube* menjadi salah satu media sosial yang mudah diakses dan praktis, sehingga menjadi situs paling populer yang ditonton oleh ribuan orang setiap hari. Video atau konten yang diunggah secara publik dapat dilihat oleh semua orang dan juga dikomentari oleh penonton, sehingga media sosial ini juga dapat digunakan untuk mengumpulkan sentimen dari orang-orang terhadap suatu topik tertentu [30].