BAB II

LANDASAN TEORI

2.1. Tinjauan Pustaka

Tinjauan pustaka ini mencakup penelitian-penelitian terdahulu yang relevan dengan topik penelitian yang diteliti. Adapun sumber penelitian sejenis yang dipilih berupa skripsi dan jurnal terkait analisis sentimen menggunakan metode *Support Vector Machine,* seperti yang terlihat pada Tabel 2.1 berikut:

Tabel 2. 1 Tinjauan Pustaka

No	Penulis	Judul	Keterangan	Kekurangan dan Kelebihan
1	(Fitriana dkk., 2021)	Analisis Sentimen Opini Terhadap Vaksin *Covid-19* pada *Twitter* Menggunakan Metode *Naïve Bayes* dan *SVM.*	Tujuan penelitian ini adalah untuk membandingkan metode klasifikasi *Naïve Bayes* dan *Support Vector Machine* dalam mengklasifikasikan data *tweet* terkait *Covid-19.* Dataset yang digunakan berjumlah 1000 *tweet* dengan kata kunci *Covid-19* dari tahun 2021. Hasil penelitian menunjukkan bahwa *SVM* memiliki performa lebih baik dengan akurasi, presisi,	Kekurangan pada penelitian ini adalah ukuran dataset yang masih relatif sedikit yaitu hanya 1000 data. kelebihan pada penelitian ini adalah melakukan perbandingan 2 metode klasifikasi

			dan *recall* sebesar 90,47%, 90,23%, dan 90,78%, sedangkan *Naïve Bayes* memiliki nilai masing-masing sebesar 88,64%, 87,32%, dan 88,13%. Waktu training *Naïve Bayes* lebih cepat dibandingkan *SVM*, yakni 8,1 detik dan 11 detik.	yaitu *SVM* dan *Naive Bayes*.
2	(Aliyya, 2020)	Analisis Sentimen Berbasis Aspek pada Ulasan Aplikasi Tokopedia Menggunakan *Support Vector Machine.*	Tujuan penelitian ini adalah untuk melakukan analisis sentimen pada ulasan aplikasi Tokopedia berbahasa Indonesia. Metode yang digunakan adalah *SVM* dengan *hyperparameter* tuning menggunakan *Gridsearch CV* dan *10-fold Cross Validation* pada *kernel Linear*, *Polynomial*, dan *RBF*. Dataset yang digunakan berjumlah 5.614 ulasan pengguna Tokopedia dari *Google Play Store*. Hasil klasifikasi	Kelebihan pada penlitian ini yaitu dilakukan pelabelan ulasan berdasarkan aspek layanan, sistem, dan kebermanfaatan memberikan wawasan yang lebih dalam tentang ulasan pelanggan, yang bisa membantu dalam

			sentimen dengan *SVM* mencapai akurasi 69,6%, sementara klasifikasi aspek memiliki akurasi 74,2%.	meningkatkan pengalaman pengguna. Kekurangan pada penelitian yakni tidak mempertimbangkan kelas sentimen netral.
3	(Natasuwarna, 2020)	Seleksi Fitur *Support Vector Machine* pada Analisis Sentimen Keberlanjutan Pembelajaran Daring.	Tujuan dari penelitian ini adalah untuk melakukan analisis sentimen terhadap komentar positif dan negatif yang muncul dari masyarakat di *Twitter* terkait pernyataan Mendikbud Republik Indonesia mengenai keberlanjutan pembelajaran daring. Dataset yang digunakan diperoleh dari *Twitter* dengan menggunakan *tools* Rapidminer sejumlah 200 data *tweet.*	Kelebihan dari penelitian ini adalah membandingkan dua metode seleksi fitur, yakni *Term Frequency (TF) dan Term Frequency-Inverse Document Frequency (TF-IDF)*, untuk mendapatkan nilai *K-Fold* pada *K-Fold Cross*

			Metode yang digunakan adalah *Support Vector Machine.* Hasil klasifikasi menggunakan metode *SVM* dengan dua jenis seleksi fitur berbeda dan variabel *k-Fold* pada *Cross Validation* menunjukkan kinerja yang memuaskan. Nilai tertinggi dalam *accuracy* dan *recall* pada kedua jenis seleksi fitur adalah 86,00% dan 85,02%. Sedangkan untuk nilai *precision*, hasil tertinggi didapatkan oleh seleksi fitur *TF-IDF* dengan nilai 87,38%.	*Validation* yang menghasilkan evaluasi tertinggi. Kekurangan dari penelitian ini adalah data yang digunakan masih sedikit, yaitu hanya 200 data *tweet*
4	(Husada and Paramita, 2021)	Analisis Sentimen Pada Maskapai Penerbangan di Platform *Twitter* Menggunakan Algoritma *Support Vector Machine* (*SVM*).	Tujuan penelitian ini adalah mengembangkan metode otomatis untuk mengklasifikasikan opini pengguna *Twitter* terhadap maskapai penerbangan menjadi tiga kelas. Metode yang digunakan adalah algoritma	Kelebihan pada penelitian ini yakni tidak hanya mengukur akurasi, tetapi juga melakukan perbandingan dengan beberapa

			multi-class Support Vector Machine (*SVM*) menggunakan *kernel RBF, Linear, Polynomial*, dan *Sigmoid*. Proses klasifikasi menggunakan pendekatan *One vs One*. Dataset terdiri dari 2000 data *tweet* berbahasa Inggris. Hasil penelitian menunjukkan bahwa algoritma *SVM* dengan *kernel RBF* dan parameter *complexity* = 10 serta *gamma* = 1 memberikan akurasi tertinggi sebesar 84,37%.	metode klasifikasi lainnya seperti *Decision Tree, Random Forest, Multinomial Naive Bayes*, dan *Gaussian Naive Bayes*, sehingga memberikan pandangan yang lebih lengkap tentang kinerja *SVM*.
5	(Haque *et al.*, 2022)	Analisis Sentimen pada *Twitter* Mengenai Ne*TF*lix Diblokir Telkom Menggunakan *Support Vector Machine.*	Tujuan dari penelitian ini adalah untuk memahami pandangan masyarakat terhadap pemanfaatan *ChatGPT*. Dataset yang digunakan berupa 18,000 *tweet*, dan analisis dilakukan dengan metode kualitatif manual dan *Latent Dirichlet allocation*	Menggunakan metode *Latent Dirichlet allocation* (*LDA*), untuk mengidentifikasi topik utama dari *tweet-tweet* yang dianalisis,

			(*LDA*) untuk mengidentifikasi topik utama. Hasil penelitian menunjukkan bahwa penggunaan *ChatGPT* menarik minat berbagai kalangan, termasuk peneliti, manajer, praktisi, penghibur, analis bisnis, dan pendidik. Mayoritas pengguna terkesan dengan kinerja *ChatGPT* dan melihat potensi besar dalam membantu tugas-tugas di berbagai domain. Namun, juga terdapat kekhawatiran etis, seperti dampak negatif pada industri pendidikan.	sehingga memudahkan dalam melihat pola dan tren pandangan.
6	(Styawati *et al.*, 2021)	Analisis Sentimen Masyarakat Terhadap Program Kartu Prakerja Pada *Twitter* Dengan Metode *Support Vector*	Tujuan dari penelitian ini adalah untuk menganalisis sentimen masyarakat Indonesia terhadap Program Kartu Prakerja di *Twitter*. Dataset yang digunakan terdiri dari data *tweet* dengan kata kunci "prakerja" yang	Kelebihan dari penelitian ini adalah melakukan perbandingan antara dua *kernel* yakni *kernel Linear* dan *kernel RBF*, sehingga

		Machine.	dikumpulkan dalam rentang waktu 22 April hingga 29 April 2021 sebanyak 2000 data. Metode yang digunakan *Support Vector Machine* (*SVM*) dengan pendekatan *One Against One.* Hasil dari klasifikasi metode *SVM* dengan membandingkan dua *kernel* didapatkan nilai akurasi pada *kernel* linear adalah 98,67% dengan *precision* 98%, *recall* 99%, dan *F1-score* 98%. Sedangkan pada *kernel RBF,* nilai *accuration* adalah 98,34% dengan *precision* 97%, *recall* 98%, dan *F1-score* 98%	memberikan informasi tentang performa dari setiap *kernel*. Selain itu penelitian ini menggunakan pendekatan *One Against One*, sehingga dapat menghasilkan hasil klasifikasi yang baik.
7	(Irfani *et al.*, 2020)	Analisis Sentimen Review Aplikasi Ruangguru Menggunakan Algoritma	Tujuan penelitian adalah melakukan analisis sentimen terhadap *review* aplikasi Ruangguru yang dikeluarkan oleh masyarakat. Dataset terdiri dari 2000 data ulasan komentar	Penelitian ini memiliki beberapa kelebihan yaitu menggunakan berbagai jenis

		Support Vector Machine.	aplikasi Ruangguru dari *Google Play Store*. Metode yang digunakan adalah *Support Vector Machine* dengan berbagai jenis *kernel* (*Linear, RBF,* dan *Polynomial*). Dilakukan *Cross Validation* pada tahap pengujian dengan nilai 2-10 *fold*. Hasil penelitian menunjukkan *kernel Linear* memiliki *accuracy* tertinggi (0.897) dan kombinasi data *training* 60% dan data *testing* 40% memberikan *accuracy* 0.900. Pengujian dengan *k-fold cross-validation* menunjukkan akurasi tertinggi pada *k-fold* 6, 9, dan 10 dengan *accuracy* 0.902.	*kernel* (*Linear, RBF,* dan *Polynomial*) pada metode *Support Vector Machine*. Selain itu dilakukan perbandingan jumlah *fold* pada *Cross Validation* untuk mencari tingkat akurasi tertinggi. Namun, kekurangan dari penelitian ini adalah rentang waktu pengumpulan data terbatas hanya pada 5 November 2019 hingga 9 November 2019.

8	(Gusti, 2022)	Analisis Sentimen Terhadap Perkuliahan Jarak Jauh Di Masa Pandemi *Covid-19* Pada Media Sosial *Twitter* Menggunakan *Textblob* Dan Algoritma *Support Vector Machine* (*SVM*).	Penelitian ini bertujuan untuk menguji kinerja metode *Support Vector Machine* dan pelabelan menggunakan *TextBlob*. Dataset yang digunakan adalah 50,001 *tweet* dari *Twitter*. Penelitian ini menerapkan metode *SVM* dengan menggunakan 3 jenis *kernel* berbeda, yaitu *Linear, RBF,* dan *Polynomial* serta metode *Grid Search CV* untuk menentukan kombinasi parameter terbaik. Hasilnya menunjukkan model *SVM* dengan *Kernel Linear* mencapai akurasi tertinggi 94.8% dan nilai *precision, recall*, serta *f1-score* masing-masing sebesar 94.5% dan 95%. Evaluasi model dengan *ROC* dan *ROC Recall Precision* menghasilkan nilai *AUC* dan *AUC Recall Precision* sebesar	Penelitian ini memiliki beberapa kelebihan, yaitu metode pelabelan menggunakan *TextBlob* berbasis *lexicon*, penggunaan berbagai *kernel* (*Linear, RBF*, dan *Polynomial*) pada algoritma *Support Vector Machine*, penerapan *GridSearch CV* untuk menentukan parameter terbaik, dan visualisasi hasil analisis sentimen dengan *Bar Chart*,

			1, menunjukkan klasifikasi yang sangat baik.	*Wordcloud,* dan *WordTree.*
9	(Bukar, *et al*., 2023)	*Text Analysis of ChatGPT as a Tool for Academic Progress or Exploitation.*	Tujuan dari penelitian ini adalah untuk menganalisis sentimen pengguna *Linkedin* terhadap pemanfaatan *ChatGPT* dalam konteks akademik. Metode *VOSviewer* digunakan untuk penambangan teks dan visualisasi data jaringan media sosial *Linkedin*. Hasilnya menunjukkan mayoritas pengguna terkesan dengan kinerja *ChatGPT* dan melihat potensi manfaatnya dalam berbagai bidang akademik. Namun, penggunaan *ChatGPT* serta pengembangannya yang berkelanjutan menimbulkan beberapa masalah etika yang signifikan yang perlu dipertimbangkan	Penelitian ini memiliki beberapa kelebihan, di antaranya adalah menjadi studi pertama yang memanfaatkan algoritma analisis teks *VOSviewer* untuk penambangan data dan visualisasi penggunaan *ChatGPT* dalam konteks pendidikan. Hasil penelitian memberikan pemahaman yang baik tentang

				pandangan civitas akademika terhadap pemanfaatan *ChatGPT* sebagai alat untuk kemajuan akademik. Penelitian ini juga memiliki kelemahan, yaitu keterbatasan waktu dan sumber daya yang tersedia, sehingga jumlah dan cakupan data yang dikumpulkan tidak terlalu besar.
10	(Siregar dkk., 2019)	*Comparison Study Of Term*	Penelitian ini bertujuan untuk membandingkan pembobotan	Kekurangan pada penelitian ini

		Weighting Optimally With SVM In Sentiment Analysis.	istilah optimal antara metode *TF-IDF*, *TF*, dan *BTO* dalam analisis sentimen menggunakan metode *SVM*. Dataset yang digunakan terdiri dari 200 *tweet* sebagai *data training* dan 10 *tweet* sebagai *data testing* dari *Twitter*. Hasil penelitian menunjukkan bahwa model *SVM* dengan teknik pembobotan *TF-IDF* memiliki kinerja yang lebih baik, sedangkan kinerja model *SVM* dengan teknik pembobotan *TF* dan *BTO* memiliki hasil yang sama.	yaitu keterbatasa dataset yang hanya menggunakan 200 data training dan 10 data *tweet* sebagai data testing. Penelitian ini juga tidak memberikan hasil pengujian dari masing-masing teknik pembobotan.

2.1.1 Literatur 1

Oleh (Fitriana dkk., 2021) dengan judul Analisis Sentimen Opini terhadap Vaksin *Covid-19* pada *Twitter* Menggunakan Metode *Naïve Bayes* dan *SVM*. Penelitian ini melakukan perbandingan metode *Naïve Bayes* dan *Support Vector Machine* dengan tujuan untuk menentukan metode klasifikasi yang paling sesuai untuk analisis sentimen terhadap opini publik tentang vaksin *Covid-19* pada media

sosial *Twitter*. Penelitian ini menggunakan data *tweet* dengan kata kunci *Covid-19* yang diambil dari tahun 2021 sebanyak 1000 data.

Berdasarkan hasil penelitian, metode *SVM* memiliki performa lebih baik dibandingkan dengan metode Naïve Bayes dalam hal akurasi, presisi, dan *recall* dengan nilai masing-masing sebesar 90,47%, 90,23%, dan 90,78%. Disisi lain, metode Naïve Bayes memiliki performa yang sedikit lebih rendah dengan nilai akurasi sebesar 88,64%, presisi sebesar 87,32%, dan *recall* sebesar 88,13%. Namun dari sisi waktu training, metode Naïve Bayes membutuhkan waktu training yang lebih cepat dengan nilai 8,1 detik dibandingkan dengan *SVM* yang memerlukan 11 detik. Hasil analisis sentimen menunjukkan bahwa pada metode Naïve Bayes, terdapat 8,76% sentimen netral, 42,92% sentimen negatif, dan 48,32% sentimen positif. Sementara pada metode *SVM*, diperoleh 10,56% sentimen netral, 41,28% sentimen negatif, dan 48,16% sentimen positif.

2.1.2 Literatur 2

Oleh (Aliyya, 2020) dengan judul Analisis Sentimen Berbasis Aspek pada Ulasan Aplikasi Tokopedia Menggunakan *Support Vector Machine*. Penelitian ini bertujuan untuk melakukan analisis sentimen pada ulasan aplikasi Tokopedia berbahasa Indonesia dan memahami informasi yang diperoleh dari setiap kelas sentimen.

Metode yang digunakan adalah *SVM* dengan teknik hyperparameter tuning menggunakan Gridsearch CV dan 10-fold Cross Validation untuk memperoleh nilai parameter yang optimal pada *kernel* Linear, Polynomial, dan RBF. Peneliti melakukan pengujian terhadap beberapa parameter, seperti parameter cost dengan

nilai 1, 10, 100, dan 1000, dan parameter gamma dengan nilai 0,01, 0,1,1,10, dan 100. Penelitian ini menggunakan *TF-IDF* sebagai teknik pembobotan kata dan *Confusion matrix* sebagai matriks evaluasi model.

Dataset yang digunakan pada penelitian ini adalah data ulasan pengguna Tokopedia dari situs Google Play Store dengan total 5.614 data, yang mencakup ulasan dari bulan April hingga Juli 2019. Dataset yang telah dikumpulkan dilakukan pembagian data menjadi data training dan testing dengan perbandingan 80:20.

Hasil penelitian menunjukkan bahwa metode *SVM* dengan *kernel* linear dan parameter c=1 memiliki akurasi tertinggi sebesar 87,5% dalam mengklasifikasikan ulasan berbahasa Indonesia tentang Tokopedia. Informasi yang diperoleh dari setiap kelas sentimen menunjukkan bahwa aspek layanan memiliki persentase ulasan negatif yang lebih tinggi dibandingkan dengan aspek sistem dan kebermanfaatan. Pelabelan sentimen menghasilkan 3816 ulasan negatif dan 1798 ulasan positif, sedangkan pelabelan aspek menghasilkan 2493 ulasan beraspek layanan, 1902 ulasan beraspek sistem, dan 1219 ulasan beraspek kebermanfaatan. Hasil dari klasifikasi sentimen dengan menggunakan *SVM* memiliki akurasi sebesar 69,6%, sedangkan klasifikasi aspek dengan *SVM* memiliki akurasi sebesar 74,2%.

2.1.3 Literatur 3

Oleh (Natasuwarna, 2020) dengan judul Seleksi Fitur *Support Vector Machine* pada Analisis Sentimen Keberlanjutan Pembelajaran Daring. Peneliti melakukan analisis sentimen terhadap komentar positif dan negatif yang muncul dari masyarakat di *Twitter* terkait pernyataan Mendikbud Republik Indonesia mengenai keberlanjutan pembelajaran daring. Metode yang digunakan dalam

penelitian ini adalah klasifikasi dengan menggunakan algoritma *Support Vector Machine* (*SVM*) dengan membandingkan dua metode seleksi fitur, yaitu Term Frequency dan *TF-IDF*, untuk memperoleh nilai k-fold pada K-Fold Cross Validation yang menghasilkan evaluasi tertinggi. Penelitian ini menggunakan 200 data *tweet* yang terdiri dari 100 komentar positif dan 100 komentar negatif menggunakan lima rasio perbandingan data latih dan data uji.

Hasil klasifikasi menggunakan metode *SVM* dengan dua jenis seleksi fitur berbeda dan variabel k-Fold pada Cross Validation menunjukkan kinerja yang memuaskan. Nilai tertinggi dalam *accuracy* dan *recall* pada kedua jenis seleksi fitur adalah 86,00% dan 85,02%. Sedangkan untuk nilai *precision*, hasil tertinggi didapatkan oleh seleksi fitur *TF-IDF* dengan nilai 87,38%. Namun. kenaikan nilai k-Fold tidak berbanding lurus dengan evaluasi, dan 8-Fold Cross Validation memberikan hasil evaluasi tertinggi pada *accuracy*, *precision*, dan *recall*.

2.1.4 Literatur 4

Oleh (Husada & Paramita, 2021) dengan judul Analisis Sentimen Pada Maskapai Penerbangan di Pla*TF*orm *Twitter* Menggunakan Algoritma *Support Vector Machine* (*SVM*). Peneltian ini bertujuan untuk mengembangkan metode otomatis yang dapat melakukan klasifikasi opini dari pengguna *Twitter* terhadap maskapai penerbangan ke dalam tiga kategori, yaitu positif, negatif, dan netral.

Metode yang digunakan dalam penelitian ini adalah pendekatan machine learning dengan memanfaatkan algoritma multi-class *Support Vector Machine* (*SVM*) dengan memanfaatkan *kernel* RBF, linear, polynomial, dan sigmoid. Proses klasifikasi multi kelas dilakukan menggunakan pendekatan One vs One dan teknik

hyperparameter tuning untuk mencari parameter terbaik. Evaluasi dilakukan dengan menggunakan metode *confusion matrix*. Dalam penelitian ini digunakan 2000 data *tweet* berbahasa Inggris dari pengguna *Twitter* terhadap maskapai penerbangan.

Hasil penelitian menunjukkan bahwa algoritma *SVM* dengan *kernel* RBF dan parameter complexity = 10 dan gamma = 1 memberikan akurasi tertinggi sebesar 84,37% dalam melakukan analisis sentimen pada data *tweet*. Pada pengujian menggunakan 10-fold cross validation dilakukan perbandingan dengan beberapa metode klasifikasi lain seperti *Decision Tree, Random Forest, Multinomial Naive Bayes*, dan *Gaussian Naive Bayes*. Didapatkan bahwa *SVM* menghasilkan rata-rata nilai akurasi tertinggi yaitu sebesar 80,41%. nilai *precision*, *recall*, dan *f1-score* terbaik juga dicapai oleh algoritma *SVM* dengan *kernel* RBF dan parameter *complexity* = 10 dan *gamma* = 1, yaitu masing-masing sebesar 84,33%, 84,67%, dan 84,50%.

2.1.5 Literatur 5

Oleh (Haque dkk., 2022) dengan judul *"I think this is the most disruptive technology" Exploring Sentiments of ChatGPT Early Adopters using Twitter Data*. Penelitian ini mengeksplorasi penggunaan *ChatGPT* dalam berbagai bidang dan mengidentifikasi permasalahan serta potensinya. Peneilitian ini bertujuan untuk memahami pandangan masyarakat terhadap pemanfaatan *ChatGPT*. Dataset yang digunakan berupa 18.000 *tweet*, dengan analisis dilakukan menggunakan metode kualitatif manual dan *Latent Dirichlet allocation* (*LDA*) untuk mengidentifikasi topik utama.

Hasil penelitian menunjukkan bahwa penggunaan *ChatGPT* menarik minat berbagai kalangan, termasuk peneliti, manajer, praktisi, penghibur, analis bisnis, dan pendidik. Mayoritas pengguna terkesan dengan kinerja *ChatGPT* dan melihat potensi besar dalam membantu tugas-tugas di berbagai domain. Namun, terdapat kekhawatiran etis, seperti dampak negatif pada industri pendidikan.

Analisis pada setiap topik menunjukkan sentimen yang beragam. Pengembangan perangkat lunak dan hiburan mendapat sentimen positif yakni 81% dan 92%. Pemrosesan Bahasa Alami mendapat sentimen positif sebanyak 83%, namun ada juga yang menyatakan keprihatinan sebanyak 14% dan beberapa *tweet* netral 3%. Pada topik Intelligence Chatbot mendapat sentimen positif 78%, namun ada juga yang mencatat dampak negatif sebanyak 20%. Dampak pada pendidikan dan bisnis mendapat sentimen positif 52%, dan sentimen negatif 32%. Pada topik Implikasi untuk mesin pencari mendapat sentimen positif 54% dan sentimen negatif 15%. Terakhir, karier dan peluang masa depan mendapat sentimen positif 75% dan ada beberapa sentimen negatif sebanyak 16%.

2.1.6 Literatur 6

Oleh (Styawati dkk., 2021) dengan judul Analisis Sentimen Masyarakat Terhadap Program Kartu Prakerja Pada *Twitter* Dengan Metode *Support Vector Machine.* Penelitian ini bertujuan untuk menganalisis sentimen masyarakat Indonesia terhadap Program Kartu Prakerja melalui pembahasan yang terjadi di *Twitter*.

Penelitian ini menggunakan metode *Support Vector Machine* dengan pendekatan One Against One dan dilakukan perbandingan dua *kernel* yaitu *kernel*

linear dan *kernel* RBF serta memanfaatkan *TF-IDF* sebagai teknik pembobotan kata. Penelitian ini menggunakan data *tweet* dengan kata kunci "prakerja" yang dikumpulkan dalam rentang waktu 22 April hingga 29 April 2021 sebanyak 2000 data.

Hasil penelitian menunjukkan bahwa dari klasifikasi menggunakan metode *SVM*, terdapat tiga kelas sentimen yang diamati, yaitu netral sebanyak 98,34%, negatif sebanyak 0,99%, dan positif sebanyak 0,66%. Hasil perbandingan dua *kernel*, yaitu linear dan RBF, menunjukkan bahwa nilai akurasi pada *kernel* linear adalah 98,67% dengan *precision* 98%, *recall* 99%, dan *F1-score* 98%. Sedangkan pada *kernel* RBF, nilai akurasi adalah 98,34% dengan *precision* 97%, *recall* 98%, dan *F1-score* 98%. Hal ini menunjukkan bahwa sentimen masyarakat terhadap program Kartu Prakerja cenderung netral sebesar 98,34%.

2.1.7 Literatur 7

Oleh (Irfani dkk., 2020) dengan judul Analisis Sentimen Review Aplikasi Ruang Guru Menggunakan Algoritma *Support Vector Machine*. Penelitian ini bertujuan untuk melakukan analisis sentimen terhadap *review* aplikasi Ruang Guru. Penelitian ini menggunakan metode *Support Vector Machine* untuk mengklasifikasikan sentimen ke dalam 2 kelas, yaitu positif dan negatif.

Pada tahap pengujian, digunakan berbagai jenis *kernel SVM*, seperti *kernel* Linear, Radial Basis Function, dan Polynomial. Selain itu, dilakukan sistem Cross Validation pada tahap pengujian dengan menggunakan nilai 2-10 fold. Penelitian ini menggunakan data ulasan komentar aplikasi Ruang Guru sebanyak 2000 data

yang dikumpulkan dalam rentang waktu 5 November 2019 – 9 November 2019 dari Google Play Store.

Hasil dari penelitian menunjukkan bahwa pengujian dengan *kernel linear* memberikan nilai *accuracy* tertinggi dibandigkan dengan *kernel* lainnya. *Kernel* Linear memberikan nilai *accuracy* sebesar 0.897. Pengujian dengan sistem k-fold cross-validation menunjukkan akurasi tertinggi pada nilai k-fold 6, 9, dan 10, dengan *accuracy* mencapai 0.902. Namun Pada k-fold 10, nilai *precision* lebih tinggi dibandingkan dengan nilai k-fold lainnya dengan nilai *precision* sebesar 0.903. Secara keseluruhan, penelitian ini menunjukkan bahwa sentimen masyarakat terhadap aplikasi Ruang Guru cenderung positif, dan nilai *accuracy* dalam penelitian ini berada di kisaran 90%.

2.1.8 Literatur 8

Oleh (Gusti, 2022) dengan judul Analisis Sentimen Terhadap Perkuliahan Jarak Jauh Di Masa Pandemi Covid-19 Pada Media Sosial *Twitter* Menggunakan Textblob dan Algoritma *Support Vector Machine*. Penelitian ini bertujuan untuk mengetahui performa dari algoritma *Support Vector Machine* dan pelabelan menggunakan TextBlob terhadap sentimen pengguna *Twitter* tentang pemberlakuan perkuliahan jarak jauh. Penelitian ini menggunakan data *tweet* sebanyak 50,001 *tweet* yang dikumpulkan dalam rentang waktu 11 Mei 2021 hingga 1 November 2021. Pada tahap pembagian data menjadi data training dan data testing, dilakukan percobaan dengan beberapa perbandingan pembagian data yakni 80:90, 70:30, 90:10, dan 60:40.

Penelitian ini menggunakan metode pelabelan Textblob berbasis Lexicon dan algoritma *Support Vector Machine* dengan memanfaatkan *kernel* Linear, RBF, dan Polynomial. Penelitian ini juga memanfaatkan metode Grid Search CV untuk menentukan kombinasi parameter *kernel* terbaik. Peneliti menggunakan nilai cost 0.1, 1, 10, dan 100. Sedangkan untuk nilai gamma, peneliti menggunakan auto. Selain itu, hasil analisis sentimen divisualisasikan dengan menggunakan Bar Chart, *Wordcloud*, dan WordTree untuk memberikan pemahaman visual yang lebih jelas. Dari klasifikasi menggunakan TextBlob, didapatkan 11,191 *tweet* positif, 12,808 *tweet* netral, dan 7,773 *tweet* negatif.

Hasil penelitian menunjukkan bahwa klasifikasi sentimen ke dalam kelas positif dan negatif menggunakan model *SVM* dengan *kernel* linear dengan nilai parameter cost 1, dan gamma auto merupakan model terbaik yang mencapai tingkat *accuracy* tertinggi sebesar 94.8%, *precision* sebesar 94.5%, *recall* sebesar 94.5%, dan *f1-score* sebesar 95%, dengan perbandingan data training dan testing sebesar 80:20. Hasil ini terbukti lebih optimal dibandingkan dengan metode klasifikasi lainnya seperti Naïve Bayes dengan akurasi 74.32%, Linear Regression dengan akurasi 82,19% , Random Forest dengan akurasi 79,57%, dan Decision Tree dengan akurasi 73,29%. Evaluasi model dengan menggunakan ROC dan ROC *Recall Precision* menghasilkan nilai AUC dan AUC *Recall Precision* sebesar 1, yang menunjukkan bahwa klasifikasi yang dilakukan sangat luar biasa.

2.1.9 Literatur 9

Oleh (Bukar dkk., 2023) dengan judul *Text Analysis of ChatGPT as a Tool for Academic Progress or Exploitation.* Penelitian ini bertujuan untuk melakukan

analisis sentimen atau pendapat para pengguna *Linkedin* terkait pemanfaatan *ChatGPT* dalam konteks akademik. Metode *VOSviewer* digunakan untuk melakukan penambangan teks dan visualisasi data jaringan media sosial *Linkedin* dalam konteks akademik dan mengidentifikasi potensi keuntungan dan risiko yang terkait dengan penggunaannya.

Hasil penelitian menunjukkan bahwa sebagian besar pengguna terpana dengan kinerja *ChatGPT* dan potensinya untuk membantu aktivitas yang berkaitan dengan ilmu data, pengembangan perangkat lunak, penelitian, penulisan manuskrip, analisis data, inisiatif bisnis, dan NLP. Namun, penelitian ini memiliki keterbatasan dalam jumlah dan cakupan data yang dikumpulkan. Implikasi dari hasil penelitian ini adalah bahwa penggunaan *ChatGPT* dapat memberikan manfaat dalam konteks akademik, namun perlu diperhatikan potensi risiko yang terkait dengan penggunaannya.

2.1.10 Literatur 10

Oleh (Siregar dkk., 2019) dengan judul *Comparison Study Of Term Weighting Optimally With SVM In Sentiment Analysis*. Penelitian ini bertujuan untuk membandingkan pembobotan istilah optimal antara metode *TF-IDF*, *TF*, dan *BTO* dalam analisis sentimen menggunakan metode *SVM*. Dataset yang digunakan terdiri dari 200 *tweet* sebagai *data training* dan 10 *tweet* sebagai *data testing* dari *Twitter*.

Pendekatan yang digunakan dalam penelitian ini adalah menganalisis pembobotan term optimal dari *TF-IDF*, *TF*, dan *BTO* menggunakan metode *SVM*.

Fitur target diekstraksi untuk memilih kumpulan data dengan prediksi sentimen positif dan negatif.

Hasil dari penelitian ini digunakan untuk memprediksi sentimen positif atau negatif pada *tweet*. Penelitian ini menghasilkan perbandingan metode pembobotan term, dan hasil terbaik dari pembobotan yang mendekati sentimen positif atau negatif akan digunakan dalam penelitian lanjutan terkait text mining. Hasil penelitian menunjukkan bahwa model *SVM* dengan teknik pembobotan *TF-IDF* memiliki kinerja yang lebih baik, sedangkan kinerja model *SVM* dengan teknik pembobotan *TF* dan *BTO* memiliki hasil yang sama.

2.1.11 Literatur 11

Oleh (Oktafiani dkk., 2023) dalam penelitian yang berjudul "Pengaruh Komposisi *Split Data* Terhadap Performa Klasifikasi Penyakit Kanker Payudara Menggunakan Algoritma *Machine Learning*," fokus utama adalah mengevaluasi pengaruh komposisi *split data training* dan *testing* terhadap performa klasifikasi penyakit kanker payudara menggunakan algoritma *machine learning*. Penelitian ini menggunakan dataset *Wisconsin Breast Cancer* (*Diagnostic*), yang berisi 569 data dengan 31 atribut. Dataset ini mengandung informasi tentang kanker payudara ganas (*malignant*) dan jinak (*benign*).

Penelitian bertujuan untuk membandingkan tiga algoritma klasifikasi, yaitu *Support Vector Machine* (*SVM*), *Random Forest*, dan *Naïve Bayes*, dalam mengelola data numerik pada metode klasifikasi data. Metode yang digunakan dalam penelitian ini adalah *Holdout validation* dan *k-fold cross validation*, yang digunakan untuk membandingkan performa algoritma-algoritma tersebut.

Hasil penelitian menunjukkan bahwa pada skema *Holdout validation*, algoritma *SVM* mencapai akurasi tertinggi sebesar 98.85% dengan perbandingan *split data* 75%:25%, sementara *Random Forest* dan *Naïve Bayes* mencapai akurasi terbaik dengan komposisi *split data* 55%:45%, masing-masing sebesar 95.85% dan 93.75%. Dalam skema *k-fold cross validation*, algoritma *SVM* tetap unggul dengan akurasi tertinggi sebesar 97.71% pada skema *k-fold* k=7. *Random Forest* mencapai akurasi terbaik sebesar 95.79% pada skema *k-fold* k=6, sementara *Naïve Bayes* mencapai akurasi terbaik sebesar 93.85% pada skema *k-fold* k=3.

Secara keseluruhan menghasilkan performa akurasi yang dihasilkan oleh skema *Holdout validation* lebih unggul dibandingkan menggunakan skema *k-fold cross validation* untuk algoritma *SVM* sebesar 98.89% pada persentase *split data* 75%:25% dan *Random Forest* dengan akurasi terbaik sebesar 95.85% pada skema 55%:45%, sedangkan untuk algoritma *Naïve Bayes* performa akurasi yang dihasilkan lebih unggul saat menggunakan skema *k-fold cross validation* yaitu menghasilkan akurasi sebsar 93.85%.

2.2. *Machine Learning*

Machine learning melibatkan pengambilan pengetahuan dari data. Ini adalah bidang penelitian yang menggabungkan elemen statistik, kecerdasan buatan, dan ilmu komputer, sering disebut sebagai analitik prediktif atau pembelajaran statistik (Müller & Guido, 2017). *Machine Learning* dapat dijadikan sebagai alat yang efektif dalam mengklasifikasikan jenis sentimen. *Machine Learning* bekerja dengan memanfaatkan data dan algoritma untuk membuat model yang dapat mengenali pola dalam kumpulan data tersebut. Model tersebut kemudian dapat

digunakan untuk memprediksi keluaran atau *output* berdasarkan pola yang ada (Nasution & Hayaty, 2019).

2.3. *Natural Language Processing*

Natual Language Processing (NLP) adalah bidang yang berkaitan dengan komputer yang memproses bahasa alami manusia, yang dapat berkisar dari tugas sederhana seperti menganalisis frekuensi kata hingga pemahaman komprehensif terhadap percakapan manusia sehingga komputer dapat memberikan respons yang bermanfaat. Teknologi berbasis NLP semakin luas digunakan dalam berbagai aplikasi, termasuk prediksi teks, pengenalan tulisan tangan, pencarian web, dan terjemahan, memainkan peran utama dalam membentuk masyarakat informasi multibahasa dengan menciptakan antarmuka manusia-mesin yang lebih alami dan meningkatkan akses ke informasi (Bird dkk., 2009). NLP bertujuan unuk memahami dan memodelkan pengetahuan terhadap bahasa, baik dari segi kata, struktur kalimat, dan konteks kata dalam kalimat. Fokus NLP tidak hanya pada pemrosesan teks, tetapi juga pada pemahaman makna dari ucapan dalam bahasa alami dan memberikan respon yang tepat, misalnya dengan melakukan tindakan tertentu atau menampilkan informasi yang relevan berdasarkan permintaan pengguna (Suciadi, 2001). Dengan menggunakan teknik dan algoritma yang kompleks, NLP memungkinkan komputer untuk berinteraksi dengan manusia menggunakan bahasa alami, membuka potensi besar dalam aplikasi seperti mesin terjemahan, chatbot, analisis sentimen, dan pemrosesan teks secara umum.

2.4. Analisis Sentimen

Analisis *sentimen* atau biasa disebut *opinion mining* merupakan cabang ilmu dari *data mining* yang melibatkan pemahaman, ekstraksi, dan pengolahan data tekstual untuk memperoleh informasi terkait *sentimen* yang terkandung dalam opini. Analisis sentimen bertujuan untuk mengidentifikasi fitur-fitur yang terdapat dalam sebuah komentar (opini, perasaan, dan emosi) yang dinyatakan dalam bentuk teks. Komentar tersebut kemudian dilakukan evaluasi untuk menentukan apakah bersifat positif, netral, atau negatif (Pang & Lee, 2008).

Analisis sentimen, berperan penting dalam menganalisis pendapat, evaluasi, sentimen, sikap dan emosi yang terkait dengan beberapa entitas seperti produk, jasa, organisasi, individu, peristiwa, topik dan atribut lainnya. Penelitian tentang analisis sentimen telah banyak dilakukan, hal ini dikarenakan analisis sentimen dapat membantu pemantauan produk untuk mengetahui tanggapan masyarakat terhadap produk tersebut (Liu, 2010).

2.5. *Twitter*

Twitter didirikan oleh Jack Dorsey pada Maret 2006 dan telah menjadi salah satu pla*TF*orm media sosial paling populer di Indonesia. *Twitter* memiliki tujuan untuk memfasilitasi interaksi sosial, memberikan kemudahan akses informasi, dan menjadi tempat bagi pengguna untuk memperoleh, berbagi, dan mendokumentasikan informasi (Fikri dkk., 2020). Melalui *Twitter*, seseorang dapat menceritakan kisah sehari-hari, menyampaikan keluhan, dan menyampaikan opini dengan cepat dan tanpa hambatan terkait tren atau topik tertentu.

Tweet merupakan elemen utama konten yang ditampilkan di profil pengguna *Twitter* yang terdiri dari teks dengan batas 140 karakter. Meskipun *tweet* dapat dilihat secara publik, pengirim *tweet* dapat membatasi pesan hanya untuk teman mereka. Pengguna *Twitter* dapat melihat dan mengikuti kicauan dari pengguna lain, yang biasa disebut sebagai pengikut (*followers*). *Twitter* dapat diakses melalui website resmi, aplikasi eksternal yang kompatibel dengan ponsel, dan melalui pesan singkat (*SMS)* di negara tertentu (*Twitter*, 2013).

2.6. *ChatGPT*

Chat Generative Pre-trained Transformer atau yang biasa dikenal *ChatGPT* adalah sebuah model kecerdasan buatan yang dikembangkan pada November 2022 dengan teknologi *deep learning* (OpenAI, 2022). *ChatGPT* dilatih pada kumpulan data percakapan yang besar, sehingga dapat menghasilkan tanggapan yang relevan dengan gaya percakapan mirip manusia dan mampu terlibat dalam berbagai konteks percakapan yang diberikan oleh pengguna (Zhai, 2022).

Dalam bidang pendidikan contohnya, *ChatGPT* memiliki potensi yang menjanjikan dalam meningkatkan proses pembelajaran (Dwivedi dkk., 2023). *ChatGPT* dapat memberikan instruksi individual kepada siswa dengan semacam bimbingan pribadi atau memberikan rancangan pembelajaran yang disesuaikan dengan kebutuhan siswa. Melalui penggunaan *ChatGPT*, siswa dapat secara langsung berinteraksi dengan chatbot yang responsif dan mampu memberikan penjelasan yang lebih mudah dipahami. Hal ini dapat membantu keterlibatan dan minat siswa terhadap proses pembelajaran serta meningkatkan hasil pembelajaran dengan memberikan pendekatan yang lebih adaptif dan personal sesuai dengan

kebutuhan unik setiap siswa (Zhai, 2022). Namun disisi lain, pemanfaatan *ChatGPT* sebagai media pembelajaran memberikan resiko atau dampak *negatif* kepada para siswa. Seperti ketergantungan siswa terhadap konten yang diberikan oleh *ChatGPT* yang dapat menghambat keterampilan kognitif seperti kemampuan berfikir kritis dan pemecahan masalah, Selain itu keakuratan dan keaslian informasi yang dihasilkan juga perlu diperhatikan, mengingat model ini dirancang berdasarkan pada data yang di-training sehingga informasi yang diberikan belum tentu benar (Kasneci dkk., 2023).

2.7. *Text Mining*

Text mining merupakan sebuah proses yang bertujuan untuk mencari informasi yang terkandung dalam koleksi teks yang besar, dengan cara mengidentifikasi pola dan hubungan menarik dalam data tekstual secara otomatis. *Text mining* melibatkan disiplin ilmu yang sangat interdisipliner, seperti bidang data, pengolahan bahasa alami, pembelajaran mesin, dan pengambilan informasi (Feldman & Sanger, 2007). Tujuan dari *text mining* adalah untuk melakukan klasterisasi, klasifikasi, pengambilan informasi, dan pencarian informasi dalam teks (Kogan & Berry, 2010). *Text mining* mengambil informasi dari berbagai sumber data teks yang dapat digunakan seperti *tweet* di *Twitter*, artikel, dan jenis data teks lainnya. Data teks ini umumnya bersifat tidak terstruktur, yang meliputi teks, video, audio, foto, dan lainnya (Fitriyah dkk., 2020). Dalam *text mining*, informasi dan pengetahuan berharga dapat diekstraksi dari data teks tersebut.

2.8. *Text Preprocessing*

Sebelum melakukan klasifikasi teks pada dokumen, langkah awal yang perlu dilakukan adalah *text preprocessing*. *Text preprocessing* juga sering dikenal sebagai tahap mempersiapkan data awal. *Text preprocessing* bertujuan untuk membersihkan dan mengubah data menjadi bentuk yang lebih terstruktur agar siap untuk dilakukan analisis lebih lanjut. Beberapa tahapan dalam *text preprocessing* (Namira, 2023) adalah sebagai berikut:

1. *Cleaning* merupakan proses menghilangkan elemen-elemen yang tidak relevan, seperti tanda baca, simbol, dan karakter khusus yang tidak memberikan informasi penting.
2. *Case Folding* merupakan suatu proses mengubah semua huruf kapital didalam teks menjadi huruf kecil. Proses ini bertujuan untuk menghindari perbedaan antara huruf besar dan huruf kecil.
3. *Normalization* merupakan proses mengubah kata yang tidak baku atau *slangword* menjadi kata yang baku sesuai dengan kamus besar bahasa Indonesia (KBBI).
4. *Stemming* merupakan proses mengubah kata menjadi bentuk yang lebih dasar. Proses ini bertujuan untuk menghindari kata-kata yang bermakna sama namun memiliki variasi yang berbeda.
5. *Removing Stopword* merupakan proses menghapus kata-kata yang tidak penting dalam data yang dianalisis. Kata yang dihapus merupakan kata-kata yang paling sering muncul dan sedikit memberikan kontribusi dalam analisis teks.

6. *Tokenizing* merupakan proses memecah kalimat berdasarkan tiap kata yang menyusunnya menjadi pecahan-pecahan terpisah yang disebut sebagai token.

2.9. *Holdout Validation*

Holdout validation adalah pendekatan paling umum dalam mengevaluasi model *Machine Learning* (*ML*). Dalam pendekatan ini, data yang tersedia dibagi menjadi dua kelompok, yaitu pelatihan (*training*) dan pengujian (*testing*) (Hastie dkk., 2008). Data pelatihan digunakan untuk melatih model klasifikasi yang berisi pengetahuan yang nantinya akan digunakan untuk memprediksi kelas sentimen yang baru. Sementara itu, data pengujian digunakan untuk mengetahui performa model klasifikasi dalam memprediksi atau mengklasifikasi data yang belum pernah dilihat sebelumnya.

Holdout validation diusulkan untuk mengatasi masalah *overfitting* yang terjadi dalam validasi *re-substitution*. Di sini, data dibagi menjadi dua bagian yang tidak tumpang tindih, dan kedua bagian ini digunakan untuk pelatihan dan pengujian secara bergantian. Bagian yang digunakan untuk pengujian adalah bagian yang disebut "*hold-out*". Hal ini dinamakan demikian karena bagian tersebut disimpan untuk pengujian, sementara model dipelajari menggunakan bagian sisa dari data (Yadav & Shukla, 2016).

Dengan demikian, *Holdout validation* dapat memiliki persentase berbeda dari data yang disimpan untuk pengujian. *Holdout validation* dapat menggunakan 20% atau bahkan 10% data yang disimpan untuk pengujian. Perlu diperhatikan bahwa dalam menggunakan *Holdout validation*, waktu yang dibutuhkan untuk

pembelajaran model relatif lebih singkat dibandingkan dengan waktu yang dibutuhkan untuk pembelajaran model menggunakan *k-fold cross validation*.

2.10. *Grid Search CV*

Grid Search CV merupakan metode yang dapat digunakan untuk mencari kombinasi parameter terbaik yang nantinya akan digunakan untuk memfasilitasi model klasifikasi yang dibangun. *Grid Search CV* akan mencari kombinasi parameter terbaik yang telah ditentukan dengan melakukan *K-fold Cross Validation* (Müller & Guido, 2017). *K-Fold Cross Validation* akan membagi data menjadi *k* kelompok yang memiliki ukuran yang sama. Selanjutnya, setiap kelompok akan secara bergantian berperan sebagai data pengujian, sementara kelompok lainnya akan berperan sebagai data latih. Jika *k-fold* dilakukan dengan *k* =10, maka dataset akan terbagi menjadi 10-*fold* dengan ukuran yang serupa, dimana 9-*fold* digunakan sebagai data pelatihan dan 1-*fold* digunakan sebagai data pengujian.

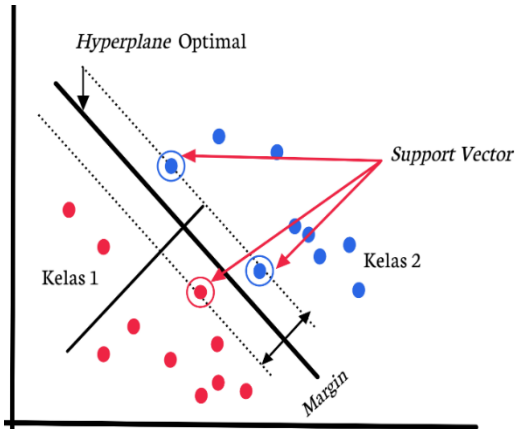
Pada setiap *fold*, *Grid Search CV* akan melakukan pengujian terhadap kombinasi parameter yang telah ditentukan. Hasil dari masing-masing *fold* akan dihitung nilai rata-ratanya untuk mendapatkan kombinasi parameter yang terbaik (Müller & Guido, 2017).

Sebagai contoh sederhana dari cara kerja dari *Grid Search CV* adalah misal terdapat parameter X = [1,2] dan Y = [3,4]. Selanjutnya *Grid Search CV* melakukan kombinasi terhadap parameter X dan Y dengan hasil [1, 3], [1, 4], [2, 3], dan [2, 4]. Kemudian dilakukan *10-fold Cross Validation*, dimana pada setiap iterasi *Grid Search CV* akan menguji setiap kombinasi parameter X dan Y serta mencatat hasil

evaluasi untuk setiap kombinasi parameter. Setelah seluruh iterasi selesai, akan didapatkan nilai rata-rata dari setiap kombinasi parameter.

2.11. *Support Vector Machine*

Support Vector Machine (*SVM*) merupakan suatu metode yang sering digunakan untuk memprediksi atau mengklasifikasikan suatu kelas berdasarkan pola yang ada. Metode ini sudah banyak digunakan dalam berbagai penelitian, termasuk dalam analisis *sentimen*. Pada dasarnya, konsep utama *SVM* adalah mencari garis pembatas (*hyperplane*) yang optimal untuk memisahkan dua kelas data yang berbeda, seperti kelas positif dan negatif (Müller & Guido, 2017)



Gambar 2. 1 Ilustrasi Support Vector Machine

Sumber (Merdekawati, 2014)

Hyperplane dapat ditemukan dengan memaksimalkan *margin* antara dua kelas data. *Margin* adalah jarak antara *hyperplane* dan data terdekat dari masing-masing kelas (Santoso, 2021). *Hyperplane* hanya dapat dipengaruhi oleh subset dari

data yang disebut sebagai *Support Vector*, yang merupakan data titik-titik yang berada tepat pada *hyperplane* atau berada dalam *margin*.

Persamaan *hyperplane* dinyatakan sebagai:

$$h(x) = w.x + b$$

Keterangan:

w = vektor yang tegak lurus dengan *hyperplane*

x = data

b = nilai bias

$h(x)$ = fungsi *hyperplane*

Persamaan *margin* dinyatakan sebagai:

$$margin = |d_{h1} - d_{h2}| = \frac{2}{||w||}$$

Keterangan:

d_{h1} = jarak *hyperplane* kelas +1

d_{h2} = jarak *hyperplane* kelas -1.

Support Vector Machine (*SVM*) umumnya digunakan untuk mengklasifikasikan data ke dalam dua kelas, jika kelas yang ingin diklasifikasikan lebih dari dua maka dibutuhkan modifikasi dengan pendekatan *SVM* multikelas. Salah satu pendekatan *SVM* multikelas yaitu *One Against One.* Dalam pendekatan ini beberapa model *SVM biner* dibangun, dimana setiap model akan membandingkan satu kelas dengan kelas lainnya. Dalam mengklasifikasikan data ke *k*-kelas, maka harus ada *k(k-1)/2* model *SVM biner* (Murphy, 2018).

Dalam Model *SVM* dengan pendekatan *One Against One* untuk klasifikasi 3 kelas (positif, negatif, dan netral), akan dibangun kombinasi pemisah biner untuk setiap pasangan kelas yang mungkin ada. Dalam kasus ini, akan dibangun 3 pemisah biner *SVM*, yaitu positif *vs*. negatif, positif *vs*. netral, dan negatif *vs*. netral. Dalam menemukan *hyperplane* terbaik yang dapat memisahkan data dan memiliki *margin* yang besar, perlu dilakukan maksimalisasi nilai *margin*. Maksimalisasi nilai *margin* dilakukan dengan menentukan jarak terbesar antara *hyperplane* dengan titik data terluar dari masing-masing kelas yang dekat dengan *hyperplane* (Murphy, 2018).

Dalam menentukan nilai *margin*, setiap model biner yang dilatih akan memiliki marginnya sendiri. *Margin* dapat ditemukan dengan menggunakan pendekatan yang serupa dengan *SVM* biner.

$$margin \quad = \frac{1}{2}\frac{(w.x_i + b)}{\|w\|} - \frac{(w.x_j + b)}{\|w\|}$$

Untuk setiap model biner, pemberian batasan pada data dari masing-masing kelas perlu dilakukan agar tidak masuk ke dalam margin, setiap kelas harus ditambahkan batasan sebagai berikut:

$$y_i(w.x_i + b) \geq 1, \ jika\ y = 1\ dan$$

$$y_i(w.x_i + b) \leq -1, \qquad jika\ y = -1$$

Yang dapat direpresentasikan menjadi,

$$y_i(w.x_i + b) \geq 1, \quad \forall\, 1 \leq I \leq n, I \in$$

Di sini, *y* mewakili label biner untuk setiap kelas dalam setiap model biner (misalnya, +1 untuk kelas yang sesuai dan -1 untuk kelas lainnya).

Untuk memaksimalkan nilai margin dengan meminimalkan *w*, kita dapat merumuskan masalah tersebut sebagai masalah optimasi *Quadratic Programming*. Persamaannya dapat dirumuskan sebagai berikut:

$$max\ margin \quad = \frac{1}{2}\|w\|^2$$

Dengan kendala:

$$y_i(w.x_i + b) \geq 1, \quad \forall\, 1 \leq I \leq n, I \in \qquad (2.1)$$

Untuk menyelesaikan persamaan diatas, dilakukan transformasi menjadi fungsi Lagrange:

$$L_p(w.b.a) \quad = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{n} \quad a_i[y_i(w.x_i + b) - 1],$$

Dimana $\alpha i \geq 0$ dengan i=1,2,3,.., *i* adalah nilai koefisien lagrange. Fungsi L_p(*primal problem*) harus dimaksimalkan terhadap α dan diminimalkan terhadap *w* dan *b*. Maka akan didapatkan kondisi sebagai berikut:

1. Kondisi 1:

$$\frac{\partial}{\partial w} L_p(w, b, a) \; = 0$$

Dari kondisi optimalitas fungsi *lagrange* diatas akan didapatkan:

$$L_p(w.b.a) \quad = \quad \frac{1}{2}\|w\|^2 - \sum_{i=1}^{n} \quad a_i y_i (w.x_i + b) + \sum_{i=1}^{n} \quad \alpha_i$$

$$\frac{\partial}{\partial w} L_p(w, b, a) \; = w - \sum_{i-1}^{n} \quad a_i y_i x_i$$

$$0 \qquad = w - \sum_{i-1}^{n} \quad a_i y_i x_i \qquad (2.\ 2)$$

$$w \qquad = \sum_{i-1}^{n} \quad a_i y_i x_i$$

2. Kondisi 2:

$$\frac{\partial}{\partial w} L_p(w, b, a) \; = \qquad\qquad 0$$

Dari kondisi optimalitas fungsi lagrange diatas akan didapatkan:

$$L_p(w.b.a) \quad = \quad \frac{1}{2}\|w\|^2 - \sum_{i=1}^{n} \quad a_i y_i (w.x_i + b) + \sum_{i=1}^{n} \quad \alpha_i$$

$$\frac{\partial}{\partial w} L_p(w, b, a) \; = \sum_{i-1}^{n} \quad a_i y_i$$

$$0 = \sum_{i-1}^{n} a_i y_i \qquad (2.3)$$

Dari kondisi optimalitas kedua, fungsi *lagrange* (*primal problem*) diubah dalam formulasi dual:

$$L_p(a) = \frac{1}{2}\left(\sum_{i-1}^{n} a_i y_i x_i\right)\left(\sum_{i-1}^{n} a_i y_i x_i\right) - \sum_{i-1}^{n} a_i y_i \left(\left(\sum_{i-1}^{n} a_i y_i x_i\right) xi + b\right) + \sum_{i-1}^{n} a_i$$

$$= \frac{1}{2}\sum_{i-1}^{n} a_i y_i a_j y_j (x_i . x_j) - \sum_{i-1}^{n} \sum_{j=1}^{n} a_i y_i a_j y_j (x_i . x_j) - b\sum_{i-1}^{n} a_i y_i + \sum_{i-1}^{n} a_i$$

$$= \sum_{i-1}^{n} a_i - \frac{1}{2}\sum_{i-1}^{n} \sum_{j=1}^{n} a_i y_i a_j y_j (x_i . x_j) \qquad (2.4)$$

Dengan kendala,

$$\sum_{i-1}^{n} a_i y = 0, \quad a_i \geq 0, i = 1,2, \ldots n$$

Untuk mencari nilai *b* (bias) digunakan persamaan berikut:

$$b = \frac{1}{2}\sum_{j-1}^{n} (y_j - \sum_{i-1}^{n} a_i y_i (x_i{}^T . x_j) \quad (2.5)$$

Dengan *dot Products* $xi.\ xj$ sering diganti dengan simbol $K(xi.\ xj)$. Maka *kernel* didefinisikan sebagai $(xi.\ xj) = (xi.\ xj)$. *Kernel* atau fungsi *kernel* digunakan untuk metransformasikan data ke dimensi yang lebih tinggi. Terdapat beberapa fungsi *kernel* yang dapat digunakan, diantaranya seperti yang ditunjukkan pada tabel 2.2 berikut.

Tabel 2. 2 Fungsi *Kernel*

No	Nama *Kernel*	Definisi Fungsi
1	*Linear*	$K(x, y) = x.y$
2	*RBF*	$K(x, y) = exp\ exp\left(-gamma\ .\sum \quad (x_i - x_j)^2\right)$
3	*Polynomial*	$K(x, y) = x.y^d$

Nilai *ai* yang diperoleh dari hasil perhitungan substitusi kendala ke persamaan LD, akan digunakan untuk menemukan nilai w. Data xi dari data latih yang memiliki nilai αi yang bernilai positif disebut support vector. Formula untuk mencari *hyperplane* terbaik adalah permasalahan *Quadratic Programming*. Setelah solusi permasalahan *Quadratic* ditemukan, maka kelas dari data tes xd dapat ditemukan berdasarkan nilai dari fungsi pemisah.

$$g(x) = sign(f(x_d))$$

$$f(x_d) = \sum_{i-1}^{ns} a_i y_i K(x_i . x_d) + b \quad (2.6)$$

2.12. Metode Evaluasi

Dalam hal klasifikasi teks, kesalahan mungkin terjadi selama proses klasifikasi. Oleh karena itu, perlu dilakukan tahap evaluasi untuk mengetahui kegunaan dan keakuratan model yang telah dibuat sebelumnya. *confusion matrix* merupakan metode evaluasi yang sering digunakan untuk mengukur model klasifikasi. *Confusion matrix* digunakan untuk membandingkan hasil klasifikasi yang didapat oleh model dengan hasil klasifikasi yang seharusnya. *Confusion matrix* pada umumnya terdiri dari dua kelas utama, yaitu positif dan negatif (Widowati, 2020). Bentuk *Confusion matrix* pada umumnya dapat dilihat pada tabel 2.3 berikut:

Tabel 2. 3 *Confusion matrix*

Confusion matrix		*Prediction Class*	
		Negatif	Positif
Actual Class	Negatif	TN	FN
	Positif	FP	TP

Sumber (Susanti, 2016)

Keterangan:

TN : *True Negatif* merupakan data sentimen negatif yang sudah diklasifikasikan dengan hasil negatif.

FN : *False Negatif* merupakan data sentimen positif yang sudah diklasifikasikan dengan hasil negatif.

FP : *False Positive* merupakan data sentimen negatif yang sudah diklasifikasikan dengan hasil positif.

TP : *True Positive* merupakan data sentimen positif yang sudah diklasifikasikan dengan hasil positif.

Confusion matrix tidak hanya dapat digunakan untuk mengevaluasi performa model klasifikasi dengan 2 kelas, namun dapat digunakan untuk mengevaluasi performa model klasifikasi dengan multikelas, seperti kelas positif, netral, dan negatif. *Confusion matrix* memainkan peran penting dalam mengevaluasi kinerja model klasifikasi multikelas. *Confusion matrix* dapat memberikan informasi mendalam tentang kemampuan model dalam memprediksi label untuk setiap kelas. Dari *confusion matrix*, terdapat beberapa metrik evaluasi yang dapat dihitung, seperti *accuracy, precission, recall,* dan *f1-score.* Metrik-metrik tersebut mengandalkan kontribusi informasi yang diberikan oleh *confusion matrix*, yang mewakili seberapa baik model mampu memprediksi label dengan benar (Grandin dkk., 2020).

1. $Accuracy = \frac{TP+FN}{TP+FP+TN+FN} \times 100\%$

Accuracy merupakan rumus memperkirakan kedekatan *True* (*Positive* dan *Negatif*) dengan nilai aktual dari kumpulan data yang lengkap.

2. *Precission* = $\frac{TP}{TP+FP} \times 100\%$

Precision merupakan rumus untuk mengukur tingkat keakuratan dalam memprediksi *True Positive* dengan jumlah keseluruhan data yang sebenarnya positif.

3. *Recall* = $\frac{TP}{TP+FN} \times 100\%$

Recall merupakan rumus untuk mengukur kemampuan dalam memprediksi *True Positive* dibandingkan dengan jumlah keseluruhan data yang diprediksi positif.

4. *F1* = $2 . \frac{Precision.recall}{Precision+recall}$

F1 score merupakan rumus yang menghitung rata-rata perbandingan antara *precision* dan *recall.*

2.13. *Wordcloud*

Wordcloud adalah alat yang digunakan untuk secara visual mempresentasikan kata-kata yang paling sering muncul dalam teks atau analisis teks. Ini berguna untuk menggambarkan metadata kata kunci dalam teks dengan cara yang bebas, yang dapat memberikan gambaran singkat tentang isu atau topik dalam penelitian. Dalam *wordcloud,* ukuran *font* kata-kata mencerminkan seberapa sering kata-kata tersebut muncul dalam teks, sehingga dengan mudah dapat mengindikasikan frekuensi kemunculan kata dalam analisis sentimen (F. Siagian & C. N. Manalu, 2020). Hal ini membantu peneliti untuk dengan cepat

mengidentifikasi kata-kata kunci yang berperan penting dalam konteks analisis sentimen mereka.

2.14. *Rapidminer*

Rapidminer merupakan perangkat lunak sumber terbuka (*open source*) yang dikembangkan oleh Dr. Markus Hofmann dari *Institute of Technology Blanchardstown*. *Rapidminer* adalah solusi yang digunakan untuk melakukan analisis data mining, text mining, serta analisis prediksi. Aplikasi ini memanfaatkan beragam teknik deskriptif dan prediksi untuk memberikan pemahaman kepada pengguna dan membantu dalam pengambilan keputusan yang terbaik. *Rapidminer* dikembangkan dalam bahasa pemrograman *Java*, sehimgga dapat digunakan di berbagai sistem operasi. *Rapidminer* telah menyediakan fasilitas analisis data yang lengkap, sehingga memudahkan pengguna untuk menganalisis data tanpa memerlukan keahlian dalam pemrograman (Srisulistiowati dkk., 2021).

2.15. *Python*

Python merupakan sebuah bahasa pemrograman berorientasi objek dengan tingkat tinggi yang dikembangkan oleh Guido van Rossum. *Python* adalah bahasa pemrograman yang telah menjadi bahasa umum atau "*lingua franca*" dalam banyak aplikasi ilmu data. *Python* menggabungkan kekuatan bahasa pemrograman umum dengan kemudahan penggunaan bahasa skrip yang berspesialisasi dalam domain seperti MATLAB atau R. *Python* memiliki berbagai perpustakaan yang mencakup berbagai bidang, termasuk pemrosesan data, visualisasi, statistik, pemrosesan bahasa alami, pemrosesan gambar, dan lainnya. Hal ini membuat *Python* menjadi

pilihan yang kuat untuk ilmu data dan berbagai aplikasi terkait (Müller & Guido, 2017). Salah satu keunggulan *python* adalah sebagai bahasa pemrograman dinamis dengan manajemen memori otomatis. Meskipun umumnya digunakan sebagai bahasa skrip, *python* memiliki penggunaan yang lebih luas yang mencakup konteks yang tidak terbatas pada skrip. Dengan keunggulan ini, *python* menjadi salah satu bahasa pemrograman yang populer dan sering digunakan dalam berbagai bidang, termasuk pengembangan aplikasi, analisis data, kecerdasan buatan, dan lainnya (Nurjanah & Insanudin, 2016).