BAB II
TINJAUAN PUSTAKA

A. Penelitian Terdahulu

Penelitian-penelitian terdahulu yang telah dilakukan oleh beberapa peneliti antara lain:

1. Penelitian yang dilakukan oleh (Jesica Krisrovina Siagian & Painem, 2024) tentang analisis sentimen masyarakat Indonesia terhadap rencana kenaikan PPN menjadi 12% di media sosial X dengan menggunakan metode *Naïve Bayes*. Studi ini menggunakan 468 dataset yang diperoleh melalui proses *crawling* data di Twitter untuk menganalisis sentimen masyarakat Indonesia terkait kenaikan PPN menjadi 12%. 326 data (77,3%) bersentimen negatif dan 106 data (22,7%) bersentimen positif sejak 1 Maret 2024- 15 Mei 2024. Penelitian ini melibatkan enam tahap utama, yaitu proses *crawling* data, pemberian label (*labeling*), *preprocessing*, pembagian data, ekstraksi fitur menggunakan *bag clasifier*, serta pengujian menggunakan *confusion matrix*. Dari hasil pengujian dan evaluasi, diperoleh akurasi sebesar 83%, *recall* sebesar 78,6%, dan presisi sebesar 68,8%.
2. Penelitian yang dilakukan oleh (Novi Fauziah & rekan-rekannya, 2024) yang berjudul pelabelan vader dalam menganalisis persepsi masyarakat terhadap kenaikan tarif PPN di Indonesia. Sentimen yang dianalisis dari media sosial Twitter menunjukkan bahwa sentimen masyarakat dominan

negatif karena banyaknya kekhawatiran serta kritikan masyarakat terkait kenaikan tarif PPN yang dirasa hanya membebani masyarakat kecil-menengah. Terdapat pula sentimen positif yang berisi dukungan serta optimisme terhadap kebijakan kenaikan tarif PPN untuk percepatan pembangunan ekonomi nasional. Adanya dominasi sentimen negatif terhadap kenaikan tarif PPN menunjukkan bahwa pentingnya bagi pemerintah untuk melihat dampak ke depan dari kenaikan tarif PPN serta survei mendalam dengan mempertimbangkan kritikan dan saran masyarakat dalam penentuan kebijakan selanjutnya. Dialog dan transparansi dari pemerintah sangat penting untuk membangun kepercayaan publik dan memastikan bahwa kenaikan PPN digunakan untuk kepentingan rakyat. Berdasarkan dari data dari 2.071 data *tweets* yang telah diolah, sentimen negatif sebesar 78%, sentimen positif memiliki persentase 6%, dan sentimen netral sebesar 16%. Sentimen negatif memiliki persentase yang lebih besar karena banyaknya kritikkan masyarakat yang tidak setuju.

3. Penelitian yang dilakukan oleh (Rahadi Rahma & teman-temannya, 2023) yang berjudul analisis sentimen pengguna Twitter menggunakan *support vector machine* pada kasus kenaikan BBM. Metode yang digunakan dalam analisis sentimen adalah *Support Vector Machine* (SVM), yang menganalisis komentar masyarakat di Twitter terkait kenaikan harga BBM. Penelitian ini memanfaatkan 258 data komentar yang diambil pada 4 September 2022, tepat sehari setelah kenaikan harga BBM. Tahap awal

penelitian mencakup *preprocessing* untuk menghapus kata-kata atau informasi yang tidak relevan. Selanjutnya, data dibagi menjadi 80% untuk pelatihan (*training*) dan 20% untuk pengujian (testing). Hasil pengujian menunjukkan tingkat akurasi sebesar 82,69%, spesifisitas 79,07%, sensitivitas 100%. Dari 52 data yang diuji, terdapat 9 komentar positif dan 43 komentar negatif, sehingga disimpulkan bahwa mayoritas masyarakat tidak setuju dengan kenaikan harga BBM.

4. Studi yang dilakukan oleh (Zidan Alhaq & koleganya, 2021) Penelitian ini membahas penerapan metode *Support Vector Machine* (SVM) untuk analisis sentimen pengguna Twitter, dengan fokus pada topik yang sering diperbincangkan, yaitu *marketplace*. Bukalapak, sebagai salah satu *marketplace* terpopuler di Indonesia, menyediakan layanan transaksi yang cepat dan aman bagi penggunanya. Ulasan dari pengguna dapat berupa sentimen positif, negatif, atau netral. Opini pengguna Bukalapak di media sosial Twitter memerlukan metode yang mampu mengidentifikasi untuk menyelesaikan permasalahan ini. Data yang diperoleh dari Twitter dilabeli dan dianalisis menggunakan metode SVM untuk mengelompokkan opini-opini pengguna. Hasil klasifikasi dengan SVM menunjukkan tingkat akurasi sebesar 93%.
5. Penelitian yang dilakukan oleh (Yuris Alkhalidi dan rekan-rekannya, 2020) mengenai analisis sentimen penghapusan ujian nasional pada twitter menggunakan *support vector machine* dan *naïve bayes* berbasis *particle swarm optimization*. Twitter digunakan sebagai platform yang membahas

tentang opini publik, hiburan dan trending topik didunia. Salah satu perbincangan pada awal tahun 2020 yakni dihapusnya Ujian Nasional (UN) oleh Kementerian Pendidikan dan Kebudayaan Republik Indonesia (Kemendikbud RI). Opini dan sentimen pengguna di *twitter* pun sangat beragam, ada yang termasuk ke dalam sentimen positif dan ada juga sentimen negatif. Untuk memilah mana yang termasuk ke dalam sentimen positif dan negatif diperlukan sebuah rangkaian proses, salah satu proses yang dapat digunakan yakni data *mining*. Pengujian dilakukan menggunakan *k-fold cros validation* untuk diperoleh nilai akurasi (*accuracy)*, tabel *confusion matrix* dan *area under curve*. Hasil pengujian diperoleh nilai akurasi 92,92% dan ACU sebesar 0,977 untuk SVM tanpa PSO. Lalu nilai akurasi 94,81% dan ACU sebesar 0,974 untuk SVM dengan PSO. Nilai akurasi 85,93% dan ACU sebesar 0,645 untuk NB tanpa PSO. Serta nilai akurasi 86,92% dan ACU sebesar 0,715.

6. Penelitian yang dilakukan oleh (Hendry Cipta Husada dan Adi Suryaputra Paramita, 2021) dengan judul analisis sentimen pada maskapai penerbangan di platform *twitter* menggunakan algoritma *support vector machine*. Perkembangan teknologi saat ini telah memberikan kemudahan bagi banyak orang dalam mendapatkan dan menyebarkan informasi di berbagai platform medial sosial. Twitter merupakan salah satu media yang kerap digunakan untuk menyampaikan opini sebagai bentuk reaksi seseorang atas satu hal. Proses analisis sentimen dilakukan dengan proses data *preprocessing*, pembobotan kata menggunakan metode TF-IDF,

penerapan algoritma, dan pembahasan atas klasifikasi. Klasifikasi opini dilakukan dengan *machine learning approach* memanfaatkan algoritma *multi-class support vector machine* (SVM). Data yang digunakan dalam penelitian ini adalah opini dalam bahasa inggris dari pengguna twitter terhadap maskapai penerbangan. Berdasarkan pengujian yang telah dilakukan, hasil klasifikasi terbaik diperoleh menggunakan SVM karena RBF pada nilai parameter C(*complexity)* = 10 dan y(*gamma)* = 1, dengan nilai akurasi sebesar 84,37% dan 80,41% Ketika menggunakan 10-*fold cros validation.*

7. Penelitian yang dilakuakan oleh (Huang, 2023) dengan judul *Sentiment analysis for social media using SVM classifier of machine learning.* Penelitian ini bertujuan untuk mengetahui seberapa baik kinerja *Support Vector Machine* (SVM) ketika diberi tugas menganalisis perasaan orang, Untuk mengevaluasi seberapa baik SVM bekerja, kami menggunakan satu kumpulan data pra-klasifikasi yang berasal dari *tweet*. Hasil dari penelitian ini adalah Metrik presisi, *recall*, dan *f-measure* digunakan untuk melakukan analisis akurasi pada hasil. Menurut penyelidikan, kumpulan data tersebut memiliki akurasi 91,8%, presisi 91,3%, dan *recall* 82,8%. Selain itu, *f1-score* dinyatakan sebesar 86,9%. Keterbatasan penelitian ini yaitu tidak ada proses normalisasi kata, sehingga masih ada *slang word* pada hasil *preprocessing*.
8. Penelitian yang dilakukan oleh (Aditiya Hermawan dan koleganya, 2023) yang berjudul implementasi *text-mining* untuk analisis sentimen pada

twitter dengan algoritma *support vector machine.* Setiap tahun, jumlah orang yang menggunakan media sosial bertambah seiring dengan jumlah orang yang menggunakan internet. Peningkatan tersebut diiringi dengan meningkatnya informasi pada internet yang tentunya informasi tersebut mempunyai nilai jika dilakukan analisis. Untuk menganalisis data dalam jumlah besar dapat menggunakan teknik *text mining. Text mining* mampu memproses untuk memperoleh informasi berkualitas tinggi dari teks. Penggunaan *text mining* menggunakan SVM dalam melakukan klasifikasi pada tweet berbahasa Indonesia mempunyai akurasi 73% berdasarkan pada 10 kali percobaan yang dilakukan dengan *keyword* dan waktu yang berbeda-beda. Kemudian nilai presisi yang didapatkan adalah 67% dan nilai *recall* yang didapatkan 54%.

Tabel 2.1 Penelitian Terdahulu

No	Peneliti	Judul	Metode	Hasil	Keterbatasan
1.	(Jesica Kristoviani Siagian dan Painem, 2024)	*Analisis sentimen Masyarakat Indonesia terhadap rencana kenaikan PPN menjadi 12% di media sosial X dengan menggunakan metode naïve bayes.*	*Naïve bayes*	Hasil yang diperoleh 468 dataset yang diperoleh melalui proses crawling data di twitter untuk menganalisis sentiment Masyarakat di Indonesia terkait rencana kenaikan PPN 12%. 326 data (77,3%) bersentimen negatif 106 data (22,7%) bersentimen positif sejak 1 maret 2024-15 mei 2024. Dari hasil pengujian diperoleh akurasi sebesar 83%, recall sebesar 78,6% dan presisi sebesar 68,8%.	Keterbatasan pada penelitian ini yaitu tidak terdapat proses normalisasi kata, yang dimana data yang berisikan kalimat tidak baku tidak diubah.
2.	(Novi Fauziah dan rekan-rekanya 2024)	*pelabelan vader dalam menganalisis presepsi Masyarakat terhadap kenaikan tarif PPN di indonesia*	*Support Vector Machine*	Hasil penelitan dari media sosial twitter menunjukkan sentimen positif yang berisi dukungan serta optimisme terhadap kebijakan kenaikan tarif PPN untuk percepatan pembangunan ekonomi nasional. Berdasarkan dari data dari 2.071 data tweets yang telah diolah, sentimen negatif sebesar 78%, sentimen positif memiliki persentase 6%, dan sentimen netral sebesar 16%. Sentimen negatif memiliki persentase yang lebih besar karena banyaknya kritikkan masyarakat yang tidak setuju	Keterbatasan pada penelitian ini yaitu penggunaan metode perluasan akronim, bahasa gaul penerjemahan kata, dan penerjemahan emoji pada tahap preprocessing.
3.	(Rahadi Rahma dan teman - temanya, 2023)	*Analisis sentiment pengguna twitter menggunakan support vector machine pada kasuk kenaikan BBM*	*Support Vector Machine*	Penelitian ini memanfaatkan 258 data komentar yang diambil pada 4 September 2022, tepat sehari setelah kenaikan harga BBM. Tahap awal penelitian mencakup *preprocessing* untuk menghapus kata-kata atau informasi yang tidak relevan. Selanjutnya, data dibagi menjadi 80% untuk pelatihan (*training*) dan 20% untuk pengujian (testing). Hasil pengujian menunjukkan tingkat akurasi sebesar 82,69%, spesifisitas 79,07%, sensitivitas 100%. Dari 52 data yang diuji, terdapat 9 komentar positif dan 43 komentar negatif.	Keterbatasan pada penelitian ini yaitu teknik TF-IDF digunakan untuk mengubah kalimat-kalimat abstrak menjadi vektor sehingga menjadi dimodelkan dengan SVM.

No	Peneliti	Judul	Metode	Hasil	Keterbatasan
4.	(Zidan Alhaq dan koleganya)	Penerapan metode *support vector machine* (SVM) untuk analisis sentiment pengguna	SMOTE dan SVM	Hasil dari penelitian ini adalah klasifikasi sentimen masyarakat terhadap pemblokiran situs judi *online* dapat dilakukan dengan menggunakan metode SVM dengan pembobotan TF-IDF dan penyetaraan data SMOTE. Klasifikasi sentimen dari teks komentar YouTube mencapai nilai akurasi sebesar 61.84% dan mencapai nilai F1-*score* 0.7590	Keterbatasan pada penelitian ini yaitu pada penelitian ini tidak dijelaskan mengenai sistem informasi, namun penelitian ini lebih menjelaskan tahap *preprocessing*
5.	(Yusri Alkhalidi, Windu Gata, Arfhan Prastyo, dan Imam Budiawan, 2020)	analisis sentimen penghapusan ujian nasional pada *twitter* menggunakan *support vector machine* dan *naïve bayes* berbasis *particle swarm optimization*	*Support Vector Machine*	Hasil Untuk memilah mana yang termasuk ke dalam sentimen positif dan negatif diperlukan sebuah rangkaian proses, salah satu proses yang dapat digunakan yakni data *mining*. Pengujian dilakukan menggunakan *k-fold cros validation* untuk diperoleh nilai akurasi (*accuracy*), tabel *confusion matrix* dan *area under curve*. Hasil pengujian diperoleh nilai akurasi 92,92% dan ACU sebesar 0,977 untuk SVM tanpa PSO. nilai akurasi 94,81% dan ACU sebesar 0,974 untuk SVM dengan PSO. Nilai akurasi 85,93% dan ACU sebesar 0,645 untuk NB tanpa PSO. Serta nilai akurasi 86,92% dan ACU sebesar 0,715.	Keterbatasan penelitian ini yaitu pada tahap *pre-processing* tidak terdapat proses *cleaning* dan normalisasi.
6.	(Hendry Cipta Husada dan Adi Syahputra Pramita)	analisis sentimen pada maskapai penerbangan di platform *twitter* menggunakan algoritma *support vector machine*	*Support vector machine*	Hasil klasifikasi opini dilakukan dengan *machine learning approach* memanfaatkan *algortimamulti-class support vector machine* (SVM). Data yang digunakan dalam penelitian ini adalah opini dalam bahasa inggris dari pengguna twitter terhadap maskapai penerbangan. Berdasarkan pengujian yang telah dilakukan, hasil klasifikasi terbaik diperoleh menggunakan SVM RBF pada nilai parameter C(*complexity*) = 10 dan y(gamma) = 1, dengan nilai akurasi sebesar 84,37% dan 80,41% Ketika menggunakan 10-*fold cros validation*.	Keterbatasan pada penelitian ini yaitu ketidakcukupan data, keakuratan prediksi pengklasifikasi dengan metode penyematan *Word2Vec* rendah sehingga makalah ini hanya mengumpulkan opini masyarakat Inggris di Twitter tentang maskapai penerbangan.

No	Peneliti	Judul	Metode	Hasil	Keterbatasan
7.	(Huang, 2023)	*Sentiment analysis for social media using SVM classifier of machine learning*	*Support Vector Machine*	Hasil dari penelitian ini adalah Metrik presisi, *recall*, dan *f-measure* digunakan untuk melakukan analisis akurasi pada hasil. Menurut penyelidikan, kumpulan data tersebut memiliki akurasi 91,8 persen, presisi 91,3 persen, dan *recall* 82,8 persen. Selain itu, nilai f1 bisa dinyatakan sebesar 86,9	Keterbatasan penelitian ini yaitu pada penelitian ini tidak ada proses normalisasi kata, sehingga masih ada *slang word* pada hasil *preprocessing*.
8.	(Aditiya Hermawan, Indrico Jowenes, Junaedi, dan Edy, 2023)	implementasi *text-mining* untuk analisis sentimen pada twitter dengan algoritma *support vector machine*.	*Support Vector Machine*	*Text mining* mampu memproses untuk memperoleh informasi berkualitas tinggi dari teks. Penggunaan *text mining* menggunakan SVM dalam melakukan klasifikasi pada tweet berbahasa Indonesia mempunyai akurasi 73% berdasarkan pada 10 kali percobaan yang dilakukan dengan *keyword* dan waktu yang berbeda-beda. Kemudian nilai presisi yang didapatkan adalah 67% dan nilai *recall* yang didapatkan 54%.	Keterbatasan penelitian ini yaitu pada penelitian ini tidak ada proses normalisasi kata, sehingga masih ada *slang word* pada hasil *preprocessing*.

B. Landasan Teori

1. *Crawling*

Crawling merupakan proses otomatis untuk mengumpulkan dan mengindeks data dari berbagai sumber seperti situs web, *database*, atau dokumen. Proses ini menggunakan perangkat lunak khusus yang disebut "*crawler*" atau "*bot*" untuk mengakses sumber data dan mengambil informasi yang dibutuhkan. Data yang dikumpulkan melalui *crawling* kemudian dapat diproses dan digunakan untuk berbagai tujuan, seperti analisis data, penelitian, atau pengembangan sistem informasi. Proses *crawling* data dimulai dengan *crawler* yang menjelajahi internet dan mengindeks serta mengumpulkan data dari berbagai sumber (Alhaq et al., 2021). Data yang dikumpulkan dapat digunakan sebagai alat untuk pengembangan sistem atau sebagai data yang biasanya digunakan oleh mesin pencarian untuk menampilkan hasil pencarian yang lebih relevan. Tujuan dari *crawling* data adalah

a. Mengumpulkan data besar dari berbagai sumber seperti situs web, *database*, atau dokumen dalam waktu singkat dan efisien.
b. Menggunakan data yang dikumpulkan untuk melakukan analisis data seperti analisis pasar, analisis perilaku pelanggan, dan lain-lain.
c. Menggunakan data yang dikumpulkan untuk melakukan penelitian seperti penelitian pasar, penelitian sosial dan lain-lain.
d. Membuat *database* yang mengandung informasi dari berbagai sumber seperti situs web, *database*, atau dokumen.

e. Memantau informasi dari berbagai sumber seperti media sosial, situs web, dan lain-lain untuk memastikan informasi yang diterima akurat dan terkini.

f. Menggunakan data yang dikumpulkan untuk membangun aplikasi seperti aplikasi pencarian, aplikasi *e-commerce*, dan lan-lain.

2. Analisis Sentimen

Analisis sentimen adalah studi komputasi yang bertujuan untuk memahami opini, sikap, dan emosi seseorang terhadap suatu topik tertentu. Hasil analisis ini biasanya diklasifikasikan sebagai sentimen positif atau negatif. Analisis sentimen melibatkan proses pendeteksian polaritas teks untuk menentukan apakah teks tersebut negatif, positif, atau netral.

3. Media Sosial X

Sosial media adalah tempat yang digunakan orang-orang untuk mengeluarkan pendapat mereka tentang berbagai topik. Pengguna sosial media di Indonesia sangat besar, hal ini mendorong munculnya data tekstual yang tidak terbatas. Salah satunya pemanfaatan data ini adalah mengetahui sentimen publik tentang kenaikan PPN 12% (Putri, 2024).

X merupakan salah satu platform media sosial yang terkenal di kalangan masyarakat umum, termasuk di Indonesia dan di seluruh dunia. Platform ini menghubungkan pengguna dengan informasi mengenai topik-topik yang relevan. *Twitter* muncul setelah kepopuleran *Facebook* sebagai platform media sosial yang mengusung konsep *microblogging*, dimana

setiap *tweet* atau cuit memiliki batasan 280 karakter. Awalnya, batasan karakter untuk setiap *tweet* adalah 140 karakter, namun jumlah ini ditingkatkan seiring berjalannya waktu. Perubahan ini mempermudah pengguna untuk mengumpulkan informasi dengan lebih efisien (Rizqi Nandadita Pamungkas et al., 2024).

Beberapa istilah yang umum digunakan di platform Twitter "X" antara lain sebagai berikut:

a. *Tweet*: Pesan atau status yang ditulis dalam kotak yang dapat berisi informasi, gambar, opini, dan rangkaian pesan lainnya. *Tweet* memiliki batasan jumlah karakter, yaitu 280 huruf.

b. *Mention*: Digunakan untuk menandai atau memanggil pengguna Twitter lain dalam sebuah *tweet* dengan menambahkan "@" diikuti dengan nama pengguna yang dimaksud.

c. *Reply*: Tanggapan atau balasan terhadap *tweet* dari pengguna lain.

d. *Retweet*: Biasa disingkat sebagai RT, digunakan untuk menunjukkan setuju dengan isi dari *tweet* pengguna lain dan membagikannya ke pengikut kita.

e. *Like*: Digunakan untuk menunjukkan bahwa pengguna menyukai *tweet* yang diunggah oleh pengguna lain.

f. *Direct Message*: Biasa disingkat sebagai DM, merupakan fitur untuk mengirim pesan secara pribadi kepada pengguna lain tanpa diketahui oleh pengikut kita.

g. *Hashtag*: Digunakan untuk menandai sebuah topik atau tema dalam sebuah *tweet* dengan menggunakan tanda "#" diikuti dengan kata kunci yang relevan. *Hashtag* membantu meningkatkan visibilitas *tweet*.

h. *Trending Topic*: Topik atau tema yang sedang populer atau banyak dibicarakan oleh pengguna Twitter karena mendapatkan perhatian yang signifikan.

4. *Google Colaboratory*

Google Colaboratory, atau yang lebih dikenal sebagai *Google Colab* merupakan sebuah alat bantu berbasis *cloud* yang disediakan secara gratis dan umum digunakan dalam kegiatan penelitian. Platform ini memungkinkan pengguna untuk menulis dan mengeksekusi kode secara langsung melalui browser tanpa perlu instalasi perangkat lunak tambahan serta beroperasi dengan menggunakan sistem penyimpanan berbasis *cloud* (Mufid, 2023). *Google collaboratory* pada dasarnya mempunyai kesamaan fungsi dengan *jupyter Notebook*, letak perbedaannya adalah *Google Collaboratory* dapat di akses secara *online* serta gratis. Beberapa fitur utama dari *google colab* adalah:

a. *Python* di *cloud*: dapat menulis dan mengeksekusi kode *python* langsung di browser web tanpa perlu mengunduh *python* atau Pustaka di komputer lokal.
b. Gratis: *Google colab* adalah layanan gratis yang disediakan oleh *Google* serta dapat menggunakannya tanpa biaya.
c. GPU gratis: *Google colab* menyediakan akses ke GPU (*Graphics Processing Unit)* secara gratis. Ini sangat berguna untuk pelatihan model mesin yang memerlukan daya komputasi yang sangat tinggi.
d. Akses ke penyimpanan *Google Drive*: dapat mengakses dan menyimpan *file* langsung di *Google Drive* yang memudahkan berbagai pekerjaan.
e. *Notebook* Interaktif: *Google Colab* menggunakan format "*notebook*" yang memungkinkan untuk menggabungkan kode, teks, gambar, dan hasil dalam satu dokumen interaktif. *Notebook* interaktif berguna untuk dokumentasi dan berbagai hasil analisis.
f. Pustaka Tersedia: *Google Colab* menyediakan banyak Pustaka umum seperti *NumPy, pandas, TensorFlow, PyTorch*, dan sebagainya yang dapat diimpor ke dalam lingkungan *colab* dengan mudah.
g. Kerja sama tim: *Google Colab* dapat berbagi *notebook colab* dengan anggota tim dan berkolaborasi dalam waktu nyata.
h. Fleksibel dan mudah digunakan: *Google Colab* adalah alat yang ramah untuk pemula memulai pemrograman *Python* dan eksplorasi data tanpa kerumitan konfigurasi lokal.

5. *Python*

Python merupakan bahasa pemrograman tingkat tinggi yang diciptakan oleh Guido van Rossum dan pertama kali dirilis pada tahun 1991. *Python* dikenal karena sintaknya yang mudah dipahami dan bersih, menjadikannya pilihan populer bagi pemula dan pengembang berpengalaman (Siagian & Painem, 2024). *Python* dilengkapi dengan pustaka standar yang sangat luas serta ekosistem pustaka pihak ketiga yang kaya, seperti *NumPy*, *Pandas*, *Matplotlib*, *TensorFlow*, *scipy, scikit learn, Theano, pytroch*, yang memudahkan pengembangan aplikasi di berbagai bidang seperti data *science*, *machine learning*, dan web *development*.

Menurut (Ma'arif, 2020) adalah bahasa pemograman tingkat tinggi yang bersifat interpreter, interaktif, objek *oriented* dan dapat berjalan di hampir semua platform. *Python* sebagai tingkat tinggi yang mudah untuk dipelajari karena sintaknya yang jelas dan juga elegan, karena sintaknya lebih menggunakan bahasa manusia daripada bahasa komputer, dan memiliki modul-modul yang dapat digunakan secara efisien. *Source code* bahasa python akan dikompilasi menjadi format *bytecode* yang akan dieksekusi. Kode python lebih lambat dieksekusi dibandingkan dengan bahasa pemrograman lain yang bersifat *low-level*. Keunggulan bahasa program python menurut (Kadarina & Ibnu Fajar, 2019):

a. Merupakan bahasa program tingkat tinggi yang mudah dipelajari karena sintaknya yang jelas dan mudah dibaca karena lebih mendekati bahasa manusia.

b. Tersedia banyak *library* yang dapat digunakan, yang kebanyakan ditulis oleh bahasa C.

c. Bahasa *python* dapat berjalan di berbagai platform tanpa harus menulis kode untuk platform tertentu.

d. Dapat digunakan untuk mengembangkan berbagai hal seperti *software*, *hardware, internet of things, web development, video game*, dan *mobile apps.*

6. *Support Vector Machine* (SVM)

Permodelan data empiris dapat menimbulkan beberapa permasalahan ketika data yang diperoleh berdimensi tinggi (ruang fitur) dan tidak seragam yang dapat melibatkan analisis dengan pendekatan *Neural Network* (NN) tradisional mengalami kesulitan dalam generalisasi dan menghasilkan model yang bisa *overfit* data. SVM dikembangkan untuk memecahkan masalah klasifikasi karena SVM memiliki kemampuan yang lebih baik dalam menggeneralisasi data bila dibandingkan dengan teknik yang sudah ada sebelumnya.

SVM merupakan sistem pembelajaran menggunakan ruang berupa fungsi-fungsi linear dalam sebuah ruangan berdimensi tinggi yang dilatih menggunakan algoritma pembelajaran berdasarkan pada teori optimasi dengan mengimplementasikan *learning bias* (Husada & Paramita, 2021). Pendekatan dengan menggunakan SVM memiliki banyak manfaat lain seperti model yang dibangun memiliki ketergantungan eksplisit pada sub

set dari *datapoints*, serta *support vector* yang membantu dalam interpretasi model.

Kelebihan SVM diantaranya efektif dalam menangani data dengan dimensi tinggi, seperti teks atau gambar, SVM hanya bergantung pada *support vectors*, sehingga membutuhkan ruang memori yang relatif kecil, melalui penggunaan kernel, SVM dapat memisahkan data yang tidak linier secara efektif (Samsudiney, 2019).

Persamaan *Support Vector Machine* (SVM) dapat dilihat pada persamaan 2.1.

$$f(x) = w.x + b \quad (2.1)$$

Keterangan :

w : Parameter yang dicari (garis yang tegak lurus antar garis dan titik *support vector*)

x : Titik data masukan *Support Vector Machine*

b : Parameter yang dicari (nilai bias)

Atau

$$f(x) = \sum_{i=1}^{M}(a_i\, y_i\, K\,(X_I, X_I) + b) \quad (2.2)$$

Keterangan:

$aiyi$: Nilai bobot setiap titik data

$K\,(x, xi)$: Fungsi kernel

b : Parameter hyperplane yang dicari (nilai bias)

Penelitian ini menggunakan kernel linear. Persamaan yaitu:

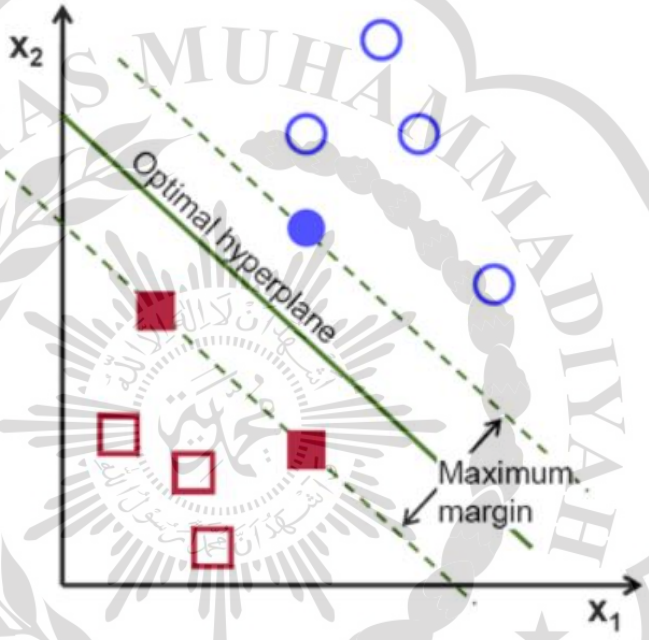
$$f(x) = w.x + b$$

$K\ (x, y)$: Nilai kernel dari data x dan y

x : Fitur data 1

y : Fitur data 2

Ilustrasi gambar metode *support vector machine* dijelaskan pada Gambar 2.1



Gambar 2. 1 Ilustrasi Metode *Support Vector Machine*

7. *Term Frequency-Inverse Document Frequency* (TF-IDF)

Term Frequency-Inverse Document Frequency (TF-IDF) adalah metode yang digunakan untuk menentukan nilai frekuensi sebuah kata di dalam sebuah dokumen atau artikel dan juga frekuensi di dalam banyak dokumen. Perhitungan ini menentukan seberapa relevan sebuah kata di dalam sebuah dokumen. TF-IDF adalah sebuah algoritma yang umumnya digunakan untuk mengolah data besar (Tanggraeni & Sitokdana, 2022).

Algoritma TF-IDF melakukan pemberian bobot pada setiap kata kunci di setiap kategori untuk mencari kemiripan kata kunci dengan kategori yang tersedia.

Sebelum melakukan pembobotan maka akan dilakukan lima tahap *text preprocessing* yaitu pemecah kalimat, *case folding, tokenizing, filtering, dan stemming,* selanjutnya dilakukan proses menghitung bobot TF-IDF, bobot *quert relevance* dan bobot *similarity* (Alhaq et al., 2021). TF-IDF pada dasarnya merupakan hasil dari perhitungan antara TF (*Term Frequency*) dan IDF (*Inverse Document Erequency*) (Sierra, 2019). Banyak cara menentukan nilai yang tepat dari kedua statistik yang ada. Dalam kasus term *frequency* tf (t, d), cara paling sederhana adalah dengan menggunakan *raw frequency* di dalam dokumen, yaitu beberapa kali term t muncul di dokumen d. jika menyatakan *raw frequency* t sebagai f (t, d), maka skema tf yang sederhana adalah tf (t,d) = f (t,d).

Rumus *Term Frequency* pada persamaan 2.3

$$TF_t = f \tag{2.3}$$

Keterangan:

TF : Frekuensi kemunculan kata dalam satu dokumen.

F : Jumlah kata pada satu dokumen.

Rumus *Invers Document Frequency* (IDF) terdapat pada persamaan 2.4 sebagai berikut:

$$IDF = Log\ \frac{N}{DF} \tag{2.4}$$

Keterangan:

N : Jumlah Dokumen

DF : Nilai TF

Rumus TF-IDF terdapat pada persamaan 2.5

$$TF.IDF = TF \ x \ IDF \quad (2.5)$$

Keterangan:

TF : Nilai TF

IDF : Nilai IDF

8. *Lexicon Based*

Lexicon based merupakan kamus atau leksikon yang digunakan untuk pemilihan kata pada data atau dokumen. Dalam implementasinya, tersedia dua kamus yaitu kamus dengan kumpulan kata yang bersentimen positif dan kamus dengan kumpulan kata yang bersentimen negatif yang digunakan untuk menjadi *wordlist* (Alvianda et al., 2019). Indonesia *Sentiment Lexicon* (Inset) merupakan komponen penting dalam pendekatan *lexicon based* yang berisi kumpulan kata atau frasa yang telah diklasifikasikan berdasarkan nilai sentimennya, baik positif, negatif, maupun netral. Pendekatan ini menggunakan Inset sebagai dasar untuk mencocokkan kata-kata dalam dokumen dengan daftar kata yang telah memiliki label sentimen. Sistem analisis sentimen akan membandingkan kata dalam teks dengan kata yang terdapat dalam inset untuk menentukan polaritas sentimen dari suatu pernyataan. Penggunaan Inset memungkinkan proses klasifikasi sentimen dilakukan secara lebih cepat dan terstruktur tanpa perlu pelatihan model. Dengan menggunakan Inset

yang akurat dan relevan, hasil analisis sentimen dapat menjadi lebih tepat dan representatif terhadap isi teks yang dianalisis (Ismail & Raden Bagus Fajriya Hakim, 2023).

Leksikal merupakan kamus yang digunakan bahasa pokok dalam metode *lexicon based.* Untuk mendeteksi klasifikasi atau sentimen, pada penelitian ini memanfatkan *library python* dengan *score polarty* < 0 adalah sentimen negatif, *score polarity* = 0 adalah sentiment netral, dan *score polarity* > 0 adalah sentimen positif. Untuk proses klasifikasi sentimen dapat dilakukan dengan persamaan 2.6 berikut:

$$S_{positive} = \sum_{iet}^{n} positive\ score_i \quad (2.6)$$

$$S_{negative} = \sum_{iet}^{n} negative\ score_i \quad (2.7)$$

$S_{positive}$ adalah bobot dari kalimat yang didapatkan melalui penjumlahan n skor polaritas kata opini positif dan $S_{negative}$ adalah bobot dari kalimat yang didapatkan melalui penjumlahan n skor polaritas kata opini negatif. Bobot pada tiap kalimat ini yang akan digunakan sebagai acuan untuk melakukan proses perbandingan. Sehingga dalam satu kalimat akan diketahui total jumlah nilai positif dan juga nilai negatif dari tiap-tiap kata penyusunannya. Dari persamaan nilai sentimen dalam satu kalimat maka diperoleh persamaan 3 untuk menentukan orientasi sentimen dengan perbandingan jumlah nilai positif, negatif dan netral.

9. Evaluasi

Tahap evaluasi yang dilakukan dengan menggunakan teknik *Confusion Matrix*. *Confusion Matrix* adalah sebuah matriks yang menunjukan bagaimana sistem klasifikasi berbasis data bekerja (Sujatmiko & Seniwati, 2019). Perhitungan dilakukan guna menentukan nilai akurasi, presisi dan *recall*. Berikut adalah algoritma yang digunakan untuk mengukur akurasi Tabel 2.2 menunjukkan ukuran evaluasi model klasifikasi.

Tabel 2. 2 Ukuran Evaluasi Model Klasifikasi

		Predicted values	
Actual Value		1 (Positive)	0 (Negative)
	1 (Positive)	TP (True Positive)	FN (false Negative)
	0 (Negative)	FP (False Positive)	TN (True Negative)

Keterangan :

a. *True Positives* (TP) : Kelas kata positif terprediksi positif
b. *True Negatives* (TN) : Kelas kata negatif terprediksi negatif
c. *False Positives* (FP) : Kelas kata negatif terprediksi positif
d. *False Negatives* (FN) : Kelas kata positif terprediksi negatif

Dalam tahap evaluasi perhitungan akan diuji dengan akurasi, presisi, *recall* dan *f1-score* yang dijelaskan sebagai berikut.:

1) Akurasi : ukuran yang menunjukkan sejauh mana hasil suatu pengukuran, prediksi, atau klasifikasi sesuai dengan nilai atau keadaan yang sebenarnya. Dalam konteks evaluasi kinerja, akurasi didefinisikan sebagai proporsi antara jumlah hasil yang benar dengan total keseluruhan pengujian atau observasi. Akurasi dapat di hitung dengan persamaan:

$$Akurasi = \frac{(TP+TN)}{(TP+TN+FP+FN)} \quad (2.8)$$

2) *Precision* (Presisi): Presisi mengukur seberapa banyak prediksi yang positif benar-benar positif dibandingkan dengan jumlah prediksi positif yang dilakukan oleh model. Dalam konteks analisis sentimen, presisi menunjukkan seberapa banyak dari semua teks yang diklasifikasikan sebagai positif yang benar-benar memiliki sentimen positif. Presisi dapat dihitung dengan persamaan :

$$Presisi = \frac{TP}{(TP+FP)} \quad (2.9)$$

3) *Recall* (*Recall*): *Recall* mengukur seberapa banyak dari semua data yang sebenarnya positif berhasil diidentifikasi oleh model sebagai positif. *Recall* menunjukkan kemampuan model untuk menangkap semua contoh positif dari data. Dalam analisis sentimen, *recall* membantu dalam menilai seberapa baik model dapat menangani kasus-kasus sentimen positif yang ada dalam dataset. *Recall* dapat dihitung dengan persamaan :

$$Recall = \frac{TP}{(TP+FN)} \quad (2.10)$$

4) *F1-Score: F1-score* adalah metrik yang menggabungkan presisi dan *recall* ke dalam satu nilai tunggal. *F1-score* adalah rata-rata harmonis dari presisi dan *recall*, dan memberikan gambaran yang lebih seimbang tentang performa model, terutama ketika ada ketidakseimbangan kelas. *F1-score* sangat berguna ketika kita membutuhkan keseimbangan antara presisi dan *recall* dan

menghindari *trade-off* antara keduanya. *F1-score* dapat dihitung dengan persamaan :

$$F1\text{-}score = \frac{TP}{TP+\frac{1}{2}(FP+FN)} \quad (2.11)$$

Atau

$$F1\text{-}score = 2 * \frac{precision*recall}{precision+recall} \quad (2.12)$$