BAB I
PENDAHULUAN

1.1 Latar Belakang

Pendidikan merupakan aspek yang amat penting dalam berbagai kehidupan, hal ini tak lepas dari besarnya peran dan dampak positif yang ditimbulkan dari majunya suatu sistem pendidikan. Dunia pendidikan merupakan upaya meningkatkan kualitas mutu sumber daya manusia dalam hal pemikiran dan keahlian. Pendidikan merupakan kunci utama bagisuatu negara untuk unggul dalam persaingan global (Yaelasari & Astuti, 2022). Pendidikan selalu berkaitan dengan kurikulum. Kurikulum merupakan suatu alat yang digunakan untuk mencapai tujuan pendidikan sehingga dapat dikatakan bahwa kurikulum merupakan acuan dalam proses penyelenggaraan pendidikan di Indonesia (Gyta dkk., 2023). Kurikulum dalam pendidikan memiliki peran yang sangat besar dalam menentukan majunya suatu pendidikan, mulai dari ranah konsep hingga aplikasi atau praktek di lapangan. Kurikulum juga memiliki peran sebagai rencana dan pengaturan mengenai isi dan bahan ajar serta pedoman cara penyelenggaraan pendidikan yang baik (Hudaidah & Ananda, 2021). Kurikulum Pendidikan di Indonesia tercatat sudah menerapkan beberapa pergantian kurikulum (Iramdan., 2019). Hal itu terkait dengan perkembangan zaman mulai dari masa pasca kemerdekaan hingga pembangunan. Pasca reformasi tahun 1998, pendidikan di Indonesia menunjukkan perubahan mendasar, dari paradigma parsial menuju holistik. Tonggak perkembangan kurikulum tahun 2004, 2006 dan 2013.

Kurikulum pendidikan indonesia telah berganti atau direvisi sekurang-kurangnya 10 kali, yaitu pada tahun 1952, 1964, 1968, 1975, 1984, 1994, 2004, 2006, 2013 (Hamami, 2021) (Intiana dkk., 2023). Kurikulum terbaru di Indonesia, kurikulum merdeka merupakan masa guru dan siswa dapat atau memiliki kebebasan dalam berpikir dan juga bebas dalam beban pikiran sehingga dapat mengembangkan potesi Pendidikan (Izza, dkk. 2020). Kurikulum merdeka belajar juga sebagai penyempurnaan dari kurikulum 2013. Salah satu keunggulan kurikulum merdeka yaitu guru dapat mengajarkan sesuai dengan capaian siswa dan siswa pun dapat mengembangkannya. Selain keunggulan ada juga kelemahannya yaitu banyaknya ketimpangan pendidikan dalam bersosialisasi sehingga membuat ketidak merataan penerapan kurikulum merdeka belajar ini (Voni, dkk. 2022). Perubahan kurikulum ini terjadi seiringan dengan perubahan sistem politik, sosial, budaya, ekonomi, dan iptek dalam kehidupan berbangsa dan bernegara (Andriani, 2020). Perubahan dan reformasi kurikulum

perlu dilakukan karena kurikulum mempunyai sifat yang dinamis agar mampu menjawab perkembangan dan tantangan zaman (Restiana dkk., 2022).

Sampai saat ini, konsep Kutikulum Merdeka telah banyak mendapat respon yang beragam dari berbagai lembaga pendidikan yang memfasilitasi pembelajaran para peserta didik, baik pada jenjang pendidikan dasar, menengah dan tinggi (Abidah dkk. 2020). Meski terdapat alasan yang kuat perihal kebijakan kurikulum merdeka Indonesia, ada beberapa hal yang menjadikan kebijakan ini menjadi pro dan kontra .Demikian penerapan kurikulum merdeka tidak lepas dari opini pro, opini kontra, dan opini netral dari masyarakat, orang tua, siswa dan guru. Adanya berbagai opini maupun komentar masyarakat terhadap kebijakan kurikulum merdeka belajar kampus merdeka pada media sosial khususnya twitter, menjadi suatu sentimen yang terjadi pada Masyarakat (Prasetyo dkk., 2021). Oleh karena itu, perlu dilakukan analisis mendalam mengenai sentimen yang terjadi pada masyarakat terkait kepuasan program kurikulum merdeka.

Penelitian ini bertujuan untuk mengklasifikasikan sentimen pengguna twitter terhadap kebijakan pemerintah mengenai kurikulum merdeka ke dalam sentimen positif dan sentimen negatif. Tujuan dari klasifikasi sentimen positif dan negative sendiri untuk memberikan informasi kepada pemerintah dan pembaca mengenai kurikulum merdeka sehingga dapat memberi masukan berdasarkan opini masyarakat. Twitter telah tumbuh menjadi situs microblogging yang populer dalam kategori aplikasi *social network*. Konten teks twitter yang menampung maksimal seratus empat puluh karakter tidak menghalangi layanan ini untuk menjadi media jejaring sosial yang handal. Hal itu dimungkinkan karena sifat pesan twitter yang bersifat singkat dan langsung sehingga memudahkan pengguna untuk menyampaikan informasi yang diinginkan (Wibowo & Winarko, 2014). Pengguna twitter dapat mengemukakan pendapatnya terhadap suatu produk atau mengomentari suatu program melalui tweet. Twitter telah menyediakan *Application Programming Interface* (API) yaitu sekumpulan fungsi atau protokol yang disediakan untuk pengguna dalamrangka mengembangkan sebuah aplikasi (Ayu, 2018). Teks twitter dapat terdiri dari beberapa bagian antara lain emoticon, URLs, RT untuk *re-tweet*, @ untuk mention pengguna lain, # untuk hashtag yang digunakan dalam penentuan topik twitter. Antar pengguna twitter yang terhubung dengan pengguna lain (follower) dapat saling melihat teks pesan yang disampaikan seorang pengguna Twitter kepada pengguna yang lain yang dikenal dengan istilah *tweet* (Sudiantoro & Zuliarso, 2018). Data teks twitter yang begitu beragam bentuk dan kandungan isinya, memiliki banyak arti jika diproses lebih lanjut, dalam konteks tersebut maka teknik Data Mining memiliki peran yang siginifikan selama data twitter tersebut bisa diperoleh dalam jumlah besar, ratusan hingga

ribuan bahkan jutaan *tweet* (Wibowo & Winarko, 2014). *Text mining* adalah salah satu teknik yang dapat digunakan untuk melakukan klasifikasi dokumen, *clustering, information extraction,* analisis sentimen dan *information retrieval* dimana *text mining* merupakan variasi dari data mining yang berusaha menemukan pola yang menarik dari sekumpulan data tekstual yang berjumlah besar. Analisis sentimen atau bisa di sebebut juga opinion mining merupakan proses memahami, mengekstrak dan mengolah data tekstual secara otomatis untuk mendapatkan informasi sentimen yang terkandung dalam suatu kalimat apakah beropini positif atau negatif (Sudiantoro & Zuliarso, 2018). Suatu algoritma untuk membagi data menjadi sebuah data training agar dapat memprediksi atau mengklasifikasikan data testing disebut sebagai supervised machine learning. Tujuan dari klasifikasi adalah mengkategorikan data ke dalam kelas label, dimana kelas label tersebut telah ditentukan oleh peneliti. Berdasarkan uraian sebelumnya maka analisis yang tepat mengenai data *text mining* menggunakan analisis sentiment maka algoritma yang akan digunakan adalah *NBC* (*Naive Bayes Classifier*) dan *SVM*. Algoritma *NBC* dan *SVM* diarasa cocok digunakan pada analisis sentimen dikarenakan algoritma ini bertujuan sebagai metode klasifikasi kedalam kategori positif dan negatif. Metode *NBC* telah banyak digunakan dalam penelitian mengenai text mining, beberapa kelebihan *NBC* diantaranya adalah algoritma sederhana tapi memiliki akurasi yang tinggi. (Ariadi & Fithriasari, 2015). Metode *NBC* telah banyak digunakan dalam penelitian mengenai *Text Mining* karena memiliki kelebihan yaitu algoritma sederhana namun memiliki akurasi yang tinggi (Rish, 2006). Teknik *SVM* didasarkan pada teori pembelajaran statistik dan telah menunjukkan hasil empiris yang menjanjikan dalam berbagai aplikasi dunia nyata, mulai dari pengenalan angka tulisan tangan hingga klasifikasi teks. *SVM* juga bekerja sangat baik untuk data dengan banyak dimensi, menghindari kesulitan karena masalah dimensi.

1.2 Rumusan Masalah

Rumusan masalah yang digunakan dalam penelitian ini adalah sebagai berikut.

1. Bagaimana karakteristik data pada *tweet* yang berkaitan dengan kebijakan Kurikulum Merdeka?
2. Bagaimana hasil klasifikasi sentimen pengguna *Twitter* terhadap kebijakan kurikulum merdeka dengan metode *Naïve Bayes Classifier*?
3. Bagaimana hasil klasifikasi sentimen pengguna *Twitter* terhadap kebijakan kurikulum merdeka dengan metode *Support Vector Machine*?
4. Bagaimana perbandingan hasil metode *Naïve Bayes Classifier* dan *Support Vector Machine*?

1.3 Tujuan

Adapun tujuan dari penelitian ini adalah sebagai berikut.

1. Mengetahui dan memahami karakteristik data pada tweet yang berkaitan dengan kebijakan Kurikulum Merdeka
2. Mengetahui dan memahami hasil klasifikasi sentimen pengguna twitter terhadap kebijakan kurikulum merdeka dengan metode *Naïve Bayes Classifier*.
3. Mengetahui dan memahami hasil klasifikasi sentimen pengguna Twitter terhadap kebijakan kurikulum merdeka dengan metode *Support Vector Machine*.
4. Mengetahui dan membandingkan hasil metode *Naïve Bayes Classifier* dan *Support Vector Machine*.

1.4 Manfaat

Penelitian ini diharapkan bisa menjadi pertimbangan pembaca dalam memilih antara metode *SVM* atau *NBC* untuk penelitian selanjutnya. Selain itu, diharapkan hasil dari penelitian ini dapat membantu pihak penyedia berita atau yang bersangkutan dalam memahami tanggapan masyarakat mengenai kurikulum merdeka untuk mempercepat proses klasifikasi tanggapan (sentimen) pengguna twitter karena telah mendapatkan model dari data training sehingga pihak penyedia berita tidak perlu mengklasifikasi ulang data secara manual. Dengan hasil penelitian ini diharapkan memberikan informasi dan referensi untuk pemerintah terkait pendapat atau sentiment melalui komentar di twitter sehingga dapat menjadi pertimbangan atau evaluasi dalam merumuskan kurikulum Pendidikan di Indonesia ke depannya.

1.5 Batasan Masalah

Penelitian ini dibatasi dengan beberapa batasan masalah sebagai berikut.

1. Data yang digunakan merupakan *tweet* yang membahas tentang kebijakan kurikulum Merdaka dengan kata kunci “kurikulum Merdaka”.
2. *Tweet* yang diambil hanya *tweet* yang menggunakan Bahasa Indonesia.
3. Peneliti tidak memperhatikan latar belakang dari pemilik akun twitter serta tidak memperhatikan waktu *tweet* dicuitkan.
4. Kernel yang digunakan pada saat klasifikasi menggunakan *SVM* adalah kernel *linier* dan kernel *Radial Basis Function* (*RBF*).

BAB II
LANDASAN TEORI

2.1 *Text Mining*

Text mining merupakan bagian dari data mining, yaitu proses untuk memperoleh suatu pengetahuan menggunakan seperangkat alat analisis dimana pengguna berinteraksi dengan sekumpulan dokumen dari waktu ke waktu. Konsep text mining biasanya digunakan dalam klasifikasi dokumen tekstual dimana dokumen-dokumen tersebut akan diklasifikasikan sesuai dengan topik dokumen tersebut(Hasri & Alita, 2022).

Dalam beberapa tahun terakhir, metode *text mining* telah digunakan untuk analisis dalam berbagai bidang diantaranya untuk pengembangan perangkat lunak dan pemasaran. Aplikasi *text mining* berasal dari cabang ilmu *data mining* yang melibatkan praproses dokumen seperti kategorisasi teks, ekstraksi informasi, dan ekstraksi kata(Feinerer et al., 2008) *Text mining* merupakan teknik yang digunakan untuk menangani permasalahan *classification*, *clustering*, *information extraction* dan *information retrieval* (Firdaus & Firdaus, 2021).

Menurut Darujati dan Gumelar (2012), *text classification* dapat dianggap proses untuk membentuk golongan (kelas-kelas) dari dokumen berdasarkan pada kelas kelompok yang sudah diketahui sebelumnya. Sistem pencarian informasi dari suatu dokumen berdasarkan *query* atau kata kunci tertentu dinamakan *Boolean retrieval model*, dimana setiap dokumen dianggap hanya sebagai kumpulan dari kata-kata (Manning, Raghavan, & Schutze, 2009). Misalnya *T* adalah himpunan semua kata yang menggambarkan isi dokumen.

$$T = \{t_1, t_2, ..., t_m\} \tag{2.1}$$

Dokumen adalah himpunan bagian dari seluruh kata, dimana D menjadi himpunan dari semua dokumen.

$$D = \{D_1, D_2, ..., D_n\} \tag{2.2}$$

Q adalah model *Boolean* dari *query* sebagai berikut.

$$Q = (W_1 \vee W_2 \vee ... \vee W_m) \wedge ... \wedge (W_i \vee W_{i+1} \vee ... \vee W_m) \tag{2.3}$$

Dimana W_i berlaku untuk D_j ketika $t_i \in D_j$. Untuk menemukan dokumen yang memenuhi Q, diberikan S_i sebagai berikut.

$$S_i = \{D_j \mid W_i\} \tag{2.4}$$

S_i adalah dokumen yang relevan dengan kata kunci. Sehingga terbentuk kumpulan dokumen yang memenuhi Q sebagai berikut.

$$(S_1 \cup S_2 \cup ... \cup S_m) \cap ... \cap (S_i \cup S_{i+1} \cup ... \cup S_m) \quad (2.5)$$

2.2 *Text Preprocessing*

Text preprocessing tidak hanya merupakan langkah penting untuk mempersiapkan korpus untuk pemodelan tetapi juga merupakan area utama yang secara langsung mempengaruhi hasil Chai, C. (2023). *Text preprocessing* merupakan langkah awal dari analisis sentimen untuk mengubah bentuk *tweet* menjadi data yang terstruktur sesuai kebutuhan. Tahapan *text preprocessing* dalam klasifikasi bertujuan untuk meningkatkan akurasi klasifikasi data. Praproses dalam *text mining* cukup rumit karena dalam Bahasa Indonesia terdapat berbagai aturan penulisan kalimat maupun pembentukan kata berimbuhan. Aturan pembentukan kata dalam Bahasa Indonesia berkaitan dengan *text preprocessing* karena hasil akhir praproses teks diharapkan mendapatkan kata dasar yang sesuai dengan Kamus Besar Bahasa Indonesia. Tahapan dalam *text preprocessing* adalah sebagai berikut.

a. *Cleansing*, yaitu tahap menghilangkan *noise* atau kata yang tidak diperlukan dalam *tweet* (Buntoro dkk, 2014). Kata yang dihilangkan adalah karakter HTML, kata kunci, ikon emosi, *hashtag* (#), *username* (@username), *url* (http://website.com), dan *e-mail* (nama@website.com).
b. *Case Folding*, merupakan proses untuk mengubah semua karakter teks menjadi huruf kecil serta menghilangkan tanda baca dan angka. Cara kerja *case folding* adalah memproses huruf alphabet dari “a” hingga “z” saja sehingga karakter selain huruf tersebut akan dihapus (Weiss, 2010).
c. *Tokenizing*, merupakan proses memecah yang semula kalimat menjadi kata-kata atau memutus urutan *string* menjadi potongan-potongan, seperti kata-kata berdasarkan tiap kata yang menyusunnya. Proses *tokenizing* dalam *Twitter* memiliki perbedaan dengan proses *tokenizing* pada teks lain . Hal ini dikarenakan adanya emoticon yang sering digunakan oleh pengguna *Twitter* dalam mengungkapkan perasaannya (Sunni & Widyantoro, 2010). Tahapan *tokenizing* dimulai dari memisah-misah bagian *tweet* yang dipisahkan dengan karakter spasi.
d. *Stemming*, yakni proses untuk mendapatkan kata dasar dengan cara menghilangkan awalan, akhiran, sisipan, dan konfiks (kombinasi dari awalan dan akhiran) (Ariadi & Fithriasari, 2015).
e. *Stopwords removal*, merupakan tahap menghapus kosakata yang bukan termasuk kata unik atau ciri pada suatu dokumen atau tidak menyampaikan pesan apapun secara signifikan pada teks atau kalimat (Dragut dkk, 2009). Kosakata yang dimaksud yaitu

seperti kata penghubung dan kata keterangan yang bukan merupakan kata unik, misalnya "dari", "akan", "seorang", dan sebagainya.

2.3 Indikator Pelabelan Metode *Lexicon* untuk Analisis Sentimen

Setelah data dibersihkan melalui tahap *text preprocessing*, maka Langkah selanjutnya adalah pelabelan menggunakan metode *lexicon* untuk menentukan sentiment positif dan negatif. Penentuan dilakukan pada data teks berupa kalimat yang memiliki kata pada kamus *lexicon* yang terdiri dari kata negatif dan positif. Kata yang teridentifikasi dalam kamus *lexicon* akan dihitung skornya sesuai dengan jumlah kata pada setiap teks atau kalimat.

$$S_{Positive} \sum_{i \in t}^{n} Positive\ Score_i$$

$$S_{Negative} \sum_{i \in t}^{n} Negative\ Score_i$$

Dimana $S_{Positive}$ 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. Dari persamaan nilai sentimen dalam satu kalimat maka diperoleh persamaan 3 untuk menentukan orientasi sentimen dengan perbandingan jumlah nilai positif, negatif dan netral.

$$Sentence_{sentiment} \begin{Bmatrix} positive\ jika\ S_{positive} > S_{Negative} \\ neutral\ jika\ S_{positive} = S_{Negative} \\ negative\ jika\ S_{positive} < S_{Negative} \end{Bmatrix}$$

Jika dalam suatu teks memiliki jumlah kata positif lebih banyak dari kata negatif, maka data teks tersebut akan dilabeli sentimen positif. Jika dalam suatu teks memiliki jumlah kata positif lebih sedikit dari kata negatif, maka data teks tersebut akan dilabeli sentimen negatif. Jika dalam suatu teks memiliki jumlah kata positif sama dengan kata negatif, maka data teks tersebut akan dilabeli sentimen netral (Ismail & Hakim, 2023).

2.4 Analisis Sentimen

Sentimen analisis atau opinion mining dapat diartikan sebagai bidang ilmu yang meneliti bagaimana menyampaikan sentimen, opini atau pendapat dan emosi yang diungkapkan di dalam teks atau kalimat. Ada beberapa topik pembahasan dalam sentimen analisis, salah satu yang paling sering diteliti yaitu klasifikasi sentimen. Topik ini berpusat pada kegiatan pengelompokan sentimen berlandaskan teks opini terhadap pembahasan masalah yang menarik (Hasri & Alita, 2022).Analisis sentimen dalam Bahasa Indonesia adalah sebuah teknik atau cara yang digunakan untuk mengidentifikasi bagaimana sebuah pendapat diekspresikan

menggunakan teks dan bagaimana sentimen tersebut bisa dikategorikan sebagai sentimen positif maupun sentimen negatif. (Djamaludin et al., 2022)

2.5 *Term Frequency Inverse Document Frequency*

Term Frequency Inverse Document Frequency (*TF-IDF*) merupakan pembobot yang dilakukan setelah ekstrasi dokumen. *TF-IDF* dilakukan agar dokumen dapat dianalisis menggunakan *Support Vector Machine. Term Frequency* (*TF*) meringkas seberapa sering suatu kata tertentu muncul dalam dokumen, sedangkan *Inverse Document Frequency* (*IDF*) menurunkan ukuran kata-kata yang sering muncul dalam dokumen. Proses metode *TF-IDF* adalah menghitung bobot dengan cara integrasi antara TF dan IDF. Langkah dalam *TF-IDF* adalah untuk menemukan jumlah kata yang kita ketahui setelah dikalikan dengan berapa banyak dokumen berita dimana suatu kata itu muncul (Ariadi D & Fithriasari K, 2015). Rumus dalam menumakan pembobot dengan *TF-IDF* adalah sebagai berikut.

$$w_{ij} = tf_{ij} \times idf \tag{2.6}$$

$$idf = \log\left(\frac{N}{df_j}\right) \tag{2.7}$$

Dimana w_{ij} adalah bobot dari kata *i* pada dokumen ke *j*, N adalah jumlah seluruh dokumen, tf_{ij} adalah jumlah kemunculan kata *i* pada dokumen *j*, dan df_j adalah jumlah dokumen *j* yang mengandung kata *i*.

2.6 *Naïve Bayes Classifier* (*NBC*)

Naïve Bayes merupakan sebuah metode pengklasifikasian dengan menggunakan probabilitas sederhana yang berakar pada Teorema *Bayes* dan memiliki asumsi ketidaktergantungan (independen) yang tinggi dari masing – masing kondisi atau kejadian (Hasri & Alita, 2022)

Teorema *Bayes* merupakan teorema yang mengacu konsep probabilitas bersyarat. Secara umum teorema *Bayes* dapat dinotasikan pada persamaan berikut.

$$P(A \mid B) = \frac{P(A)P(B \mid A)}{P(B)} \tag{2.8}$$

Dimana:

P(A|B) : Peluang terjadinya A jika kejadian B terjadi

P(A) : Peluang terjadinya A

P(B|A) : Peluang terjadinya B jika kejadian A terjadi

P(B) : Peluang terjadinya B

Metode *Naïve Bayes Classification* (*NBC*) merupakan salah satu metode yang dapat mengklasifikasikan teks. Kelebihan *NBC* adalah sederhana tetapi memiliki akurasi yang tinggi (Tan dkk, 2008). Dalam algoritma *NBC* setiap dokumen direpresentasikan dengan pasangan atribut "x_1, x_2, x_3, ... x_{i+1}, …, x_n" dimana x_1 adalah kata pertama, x_2 adalah kata kedua dan seterusnya. Sedangkan himpunan kategori topik yang dibahas disimbolkan dengan Y. Pada saat klasifikasi, algoritma akan mencari probabilitas tertinggi dari semua kategori data yang diujikan (Y_{MAP}). Adapun persamaan Y_{MAP} adalah sebagai berikut.

$$Y_{MAP} = \arg\max_{y_j \in Y} P(y_j) \prod_i P(x_i \mid y_j) \tag{2.9}$$

Nilai $P(y_j)$ dihitung pada saat *training*, didapatkan dari persamaan berikut.

$$P(y_j) = \frac{|doc\ j|}{|training|} \tag{2.10}$$

Dimana |*doc j*| merupakan jumlah *tweet* yang memiliki kategori *j* dalam *training*. Sedangkan |*training*| merupakan jumlah *tweet* dalam contoh yang digunakan untuk *training*. Untuk setiap probabilitas kata x_i untuk setiap kategori $P(x_i \mid y_j)$, dihitung pada saat *training*.

$$P(x_i \mid y_j) = \frac{z_i + 1}{|z + kosakata|} \tag{2.11}$$

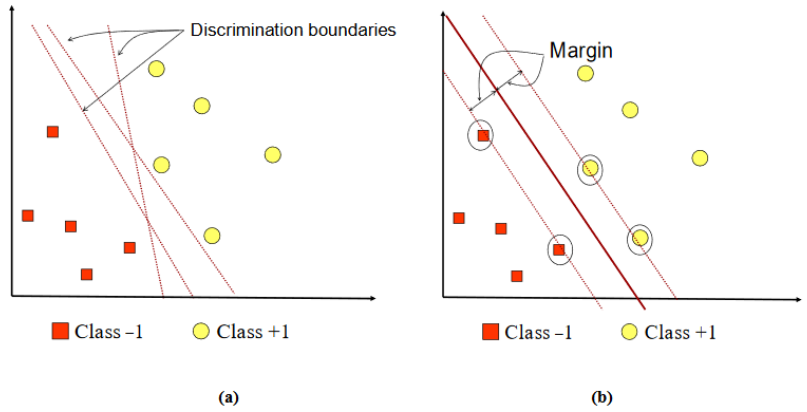
Dimana z_i adalah jumlah kemunculan kata y_i dalam *tweet* yang berkategori y_j, sedangkan *z* adalah banyaknya seluruh kata dalam *tweet* dengan kategori y_j dan |kosakata| adalah banyaknya kata dalam data *training*.

2.7 *Support Vector Machine* (*SVM*)

Support Vector Machine (*SVM*) adalah metode yang mempelajari area yang memisahkan antar kategori dalam sebuah observasi. Dalam terminology *SVM*, kita membahas jarak atau margin antar kategori. Setiap kategori memiliki observasi dimana nilai variabel targetnya sama. *SVM* juga dikenal sebagai sistem pembelajaran yang menggunakan hipotesis fungsi *linear* dalam ruang dimensi tinggi dan dilatih dengan algoritma berdasarkan toeri optimasi dengan menerapkan *learning bias* yang berasal dari teori statistik. Tujuan dari metode ini adalah membangun pemisah optimum yang disebut *OSH* (*Optimal Separating Hyperplane*) sehingga dapat digunakan untuk klasifikasi (Ariadi D & Fithriasari K, 2015).

2.7.1 Metode *SVM* pada *Linearly Separable Data*

Konsep dari *SVM* adalah berusaha menemukan *hyperplane* yang optimum pada *input space*. Fungsi dari *hyperline* itu digunakan sebagai pemisah dua buah kelas pada *input space*. Kelas biasanya sering -1 dan +1. *Hyperplane* dalam *SVM* dapat diilustrasikan pada Gambar 2.1 berikut.



Gambar 2. 1 Ilustrasi *Hyperplane* pada *SVM*

Gambar 2.1 (a) menunjukkan alternative garis pemisah antara kedua kelas (*discriminant boundaries*). Garis pemisah yang terbaik adalah yang memiliki *margin hyperplane* maksimum. *Margin* adalah jarak antara *hyperplane* dengan *pattern* terdekat pada masing-masing kelas. *Pattern* yang paling dekat disebut sebagai *support vector*. Pada Gambar 2.7 (b), *pattern* yang dilingkari adalah *support vector* untuk tiap kelas. Sedangkan, garis tebal dalam Gambar 2.7 (b) adalah *hyperplane* terbaik karena berada di tengah-tengah kedua kelas. Proses mencari letak *hyperplane* ini adalah inti dari metode *SVM*.

Data yang akan diolah dengan metode *SVM* akan dinotasikan sebagai $\vec{x}_i \in R^d$ dan untuk label masing-masing data akan dinotasikan sebagai $y_i \in \{-1,+1\}$, $i = 1,2,3,\ldots,l$ dimana l adalah banyak data yang digunakan. Sehingga fungsi *hyperplane* yang memisahkan kelas -1 dan +1 didefinisikan sebagai $h(\vec{x}_i) = 0$ atau dapat dituliskan dalam persamaan sebagai berikut.

$$\vec{w}.\vec{x} + b = 0 \tag{2.12}$$

Pattern $\vec{x}_i$ akan masuk ke dalam kelas negatif jika memenuhi pertidaksamaan $\vec{w}.\vec{x} + b \leq -1$ dan akan masuk ke dalam kelas positif jika memenuhi pertidaksamaan $\vec{w}.\vec{x} + b \leq +1$. Jika kedua pertidaksamaan tersebut dikalikan dengan y_i maka akan menjadi pertidaksamaan sebagai berikut.

$$y_i(\vec{w}.\vec{x} + b) \geq 1, \quad i = 1,2,\ldots,l \tag{2.13}$$

Jarak antara *hyperplane* dengan titik sampel terdekatnya adalah $\frac{2}{\| w\|}$. Maka jarak tersebut dapat dioptimumkan dengan cara $\max \frac{1}{\| w\|}$ atau $\min \frac{1}{2} \| w\|^2$. Kemudian didapatkan persamaan *Lagrange Multiplier* sebagai berikut.

$$L(\vec{w}, b, \alpha) = \frac{1}{2} \| w\|^2 - \sum_{i=1}^{l} \alpha_i \{ y_i [\vec{w}.\vec{x} + b] - 1 \} \quad (2.14)$$

Dimana α_i adalah *Lagrange Multiplier* yang nilainya nol atau positif $(\alpha_i \geq 0)$. Persamaan (2.12) dapat dioptimumkan dengan cara meminimalkan *L* terhadap $\vec{w}$ dan *b*, kemudian memaksimalkan *L* terhadap α_i. Jika diketahui bahwa pada titik optimum gradien *L*=0, maka didapatkan persamaan berikut.

$$\max \sum_{i=1}^{l} a_i - \frac{1}{2} \sum_{i-1}^{l} \sum_{j=1}^{l} a_i a_j y_i y_j \vec{x}_i \vec{x}_j \quad (2.15)$$

2.7.2 Fungsi Kernel pada *SVM*

Pada kasus nyata, sangat jarang dijumpai masalah yang bersifat *linier separable*. Sebagai metode yang dikembangkan untuk kasus linier, *SVM* membutuhkan sebuah fungsi yang mampu membuat pemisah yang tidak linier. Fungsi yang sering digunakan untuk mengatasi hal tersebut adalah fungsi kernel. Pada kasus nonlinier *SVM*, data $\vec{x}$ terlebih dahulu dipetakan oleh fungsi $\Phi(\vec{x})$ ke dalam ruang vektor yang berdimensi lebih tinggi. Setelah mendapat ruang vektor yang baru, kemudian *hyperplane* bisa dikonstruksikan untuk memisahkan kedua kelas sentimen yaitu positif maupun sentimen negatif didalam *tweet* yang sudah ditentukan secara manual.

Pemetaan *hyperplane* pada dimensi yang lebih tinggi ditunjukkan oleh notasi $\Phi : \Re^d \rightarrow \Re^q ; d < q$. Pemetaan yang dilakukan tidak mengubah karakteristik atau dalam kata lain tetap menjaga tipologi data. Artinya, dua data yang berjarak dekat pada *input space* akan tetap berjarak dekat juga pada *feature space*. Sebaliknya, dua data yang berjarak jauh pada *input space* akan tetap berjarak jauh pada *feature space*. Setelah melakukan pemetaan, langkah selanjutnya dalam proses *SVM* adalah menemukan titik-titik *support vector*. Untuk menemukan titik-titik *support vector*, digunakan *dot product* dari data yang sudah ditransformasi pada ruang yang berdimensi lebih tinggi, yaitu $\Phi(\vec{x}_i).\Phi(\vec{x}_j)$. Akan tetapi, transformasi Φ tidak da-pat diketahui dan sangat sulit dipahami sehingga perhitungan *dot*

product tersebut dapat diganti dengan fungsi kernel $K(\vec{x}_i, \vec{x}_j)$ yang mendefinisikan secara implisit transformasi Φ. Gunn (1998) merumuskan fungsi kernel sebagai berikut.

$$K(\vec{x}_i, \vec{x}_j) = \Phi(\vec{x}_i).\Phi(\vec{x}_j) \tag{2.16}$$

Dengan menggunakan fungsi kernel proses menentukan *support vector* menjadi lebih mudah karena cukup dengan mengetahui fungsi kernel yang dipakai, tanpa perlu mengetahui wujud dari fungsi nonlinier Φ. Beberapa fungsi kernel yang biasa digunakan ditunjukkan pada tabel berikut.

Tabel 2. 1 Fungsi *Kernel* pada *SVM*

Jenis Kernel	Fungsi
Polynomial	$K(\vec{x}_i, \vec{x}_j) = ((\vec{x}_i, \vec{x}_j)+1)^p$, *di mana* $p = 1, \ldots$
Radial Basis Function (RBF)	$K(\vec{x}_i, \vec{x}_j) = \exp\left(-\frac{\| \vec{x}_i - \vec{x}_j \|^2}{2\gamma^2}\right)$
Sigmoid	$K(\vec{x}_i, \vec{x}_j) = \tan(\alpha.\vec{x}_i \vec{x}_j + \beta)$

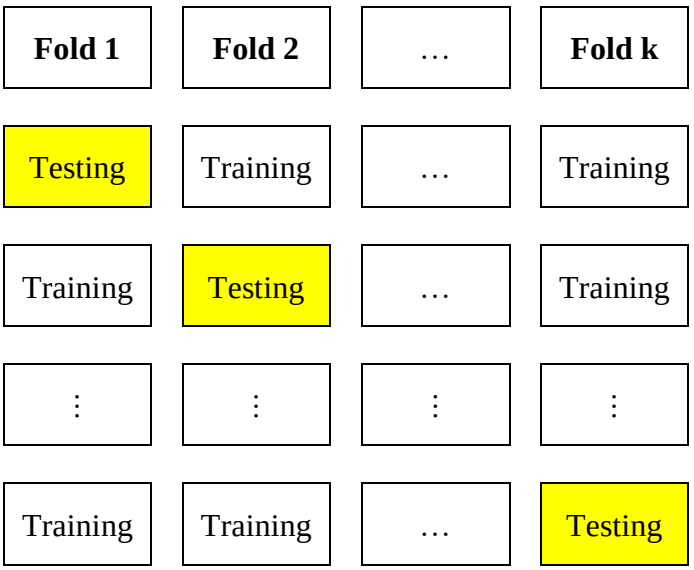
Hasil klasifikasi dari $\vec{x}$ diperoleh dari persamaan berikut.

$$\begin{aligned} f[\Phi(\vec{x})] &= \vec{w}\,\Phi(\vec{x}) + b \\ &= \sum_{i=1,SVs}^{l} \alpha_i y_i \Phi(\vec{x}).\Phi(\vec{x}_i) + b \\ &= \sum_{i=1,SVs}^{l} \alpha_i y_i K(\vec{x}_i, \vec{x}_j) + b \end{aligned}$$

Persamaan di atas merupakan subset dari *training set* yang terpilih sebagai *support vector*, artinya $\vec{x}_i$ berkorespondensi pada $\alpha_i \geq 0$. Metode *SVM* dengan kernel RBF ditambahkan parameter *C* dan γ, pada kernel *polynomial* ditambahkan parameter *C*, γ dan *ρ* sedangkan untuk kernel linier tidak ditambah dengan parameter apapun.

2.8 *K-Folds Cross Validation*

K-fold cross validation adalah salah satu metode yang digunakan untuk memastikan bahwa setiap kelas dalam data telah terwakili oleh data *training* dan *testing*. Metode ini banyak digunakan peneliti karena dapat mengurangi bias yang terjadi dalam pengambilan sampel. *K-fold cross validation* secara berulang-ulang membagi data menjadi partisi-partisi dimana setiap data mendapat kesempatan menjadi data *training* dan *testing* (Witten dkk, 2011). K merupakan besar angka partisi data yang digunakan untuk pembagian *training* dan *testing*. Berikut merupakan gambaran pembagian data menggunakan *K-fold cross validation*.



Gambar 2. 2 Ilustrasi Pembagian Data *Training* dan *Testing*

2.9 Pengukuran Performa Klasifikasi

Pengukuran performa dilakukan untuk melihat hasil yang didapatkan dari klasifikasi. Terdapat beberapa cara untuk me-ngukur performa, beberapa cara yang sering digunakan adalah dengan menghitung akurasi, *recall*, dan *precision*. Akurasi merupakan persentase dari total dokumen yang teridentifikasi secara tepat dalam proses klasifikasi. *Recall* mengindikasikan sebagian kecil dari dokumen yang relevan diambil. *Precision* mengkuantifikasi fraksi dokumen diambil yang sebenarnya relevan, dalam contoh milik kelas sasaran (Muller & Guido, 2016). Peluang ketepatan masing-masing klasifikasi dapat dilihat pada tabel berikut.

Tabel 2. 2 Perhitungan Ketepatan Klasifikasi

Observasi	Prediksi	
	Positive	*Negative*
Positive	*True Positive (TP)*	*False Negative (FN)*
Negative	*False Positive (FP)*	*True Negative (TN)*

Akurasi, *precision* dan *recall* dapat dihitung dengan persamaan berikut.

$$akurasi = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.17}$$

$$precision = \frac{TP}{TP + FP} \tag{2.18}$$

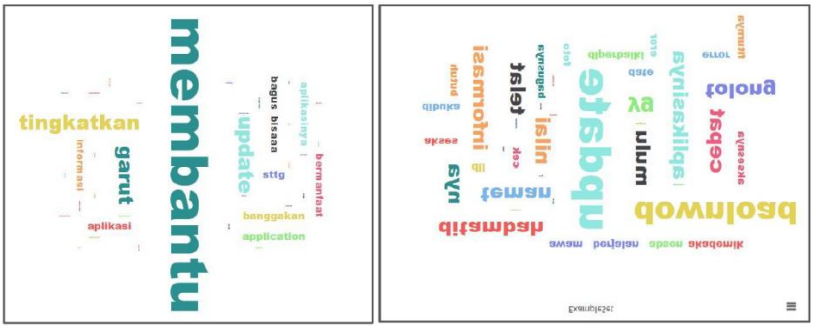
$$recall = \frac{TP}{TP + FN} \tag{2.19}$$

Untuk mengukur performa keseluruhan klasifikasi dapat menggunakan *F-measure* dengan perhitungan berikut.

$$F = 2 \times \frac{recall \times precision}{recall + precision} \tag{2.21}$$

2.10 *Word Cloud*

Word Cloud adalah gambaran visual berdasarkan frekuensi kemunculan kata-kata pada suatu kumpulan teks, dimana ukuran huruf menentukan frekuensi kemunculan sebuah kata yang artinya semakin besar ukuran huruf maka semakin besar kemunculan kata tersebut dan sebaliknya, semakin kecil huruf maka semakin kecil frekuensi kemunculan kata (Tri & Algoritma, n.d.). Berikut merupakan contoh dari visualisasi dokumen teks dengan *word cloud*.



Gambar 2. 3 Visualisasi *Word Cloud* (Sumber: (Tri & Algoritma, n.d.))