BAB II
TINJAUAN PUSTAKA DAN DASAR TEORI

2.1 Tinjauan Pustaka

Penulis menggunakan beberapa referensi sebagai bahan perbandingan dengan penelitian yang akan penulis lakukan.

Risky Maulana pada tahun (2016) dalam penelitian terhadap tokoh-tokoh publik Indonesia menggunkan Twitter Streaming API dan metode *support vector machine.* Penelitian ini memanfaatkan Pustaka lib SVM sebagai salah satu *machine learning untuk text classification,* algoritma poter pada proses stemming dan metode *term frequency* untuk pembobotan. Penelitian dari data 1.400 tweet dan 200 data uji menghasilkan akurasi sebesar 79,5%. Data yang diterima ditampilkan dan divisualisasi berupa pie chart dengan presentase hasil klasifikasi positif, negatif, netral.

Helda Ludya Safitri pada tahun (2020) dalam penelitian terhadap tindakan pemerintah Indonesia pada kasus covid-19 dengan menggunakan metode *Support Vector Machine* (SVM) dengan Kernel *Radial Basis Function* (RBF). Data yang digunakan sebanyak 600 *tweet* yang diperoleh dari hasil scraping menggunakan *twitterscraper.* Hasil dari penelitian ini mendapat tingkat akurasi pelatihan sebesar 77% dalam melakukan klasifikasi sentimen positif, negatif, dan netral. Dari hasil klasifikasi data, diperoleh sebagian besar *tweet* terdiri dari sentimen negatif.

Samsir dkk (2021) dalam penelitian terhadap penerapan algoritma *Naïve Bayes* sebagai klasifikasi sentimen pada pembelajaran daring pada masa pandemi covid-19. Hasil penelitian menunjukkan terdapat 30% sentimen positif, 69%

sentimen negatif, dan 1% netral. Tingginya sentimen negatif dihasilkan karena ketidakpuasaan masyarakat terhadap pembelajaran daring.

Rizqi alfiyati (2021) dalam penelitian terhadap opini masyarakat dari media sosial twitter pada topik kenaikan iuran bpjs. Data yang digunakan berupa data tweet berjumlah 2250 yang dibagi menjadi dua yaitu 1800 data latih dan 450 data uji dengan menggunakan metode *Support Vector Machine*. Hasil penelitian tersebut mendapat nilai akurasi 85%.

Primandani Arsi dan Retno Waluyo (2021) dalam penelitian terhadap wacana pemindahan ibu kota Indonesia. Klasifikasi dilakukan dengan cara mengklasifikasikan dalam 2 kelas yakni positif dan negatif. Berdasarkan hasil pengujian yang dilakukan terhadap *tweet* sentimen pemindahan ibu kota dari media sosial sebanyak 1.236 *tweets* (404 positif dan 832 negatif) menggunakan *Support Vector Machine* diperoleh akurasi sebesar =96,68%, *precision*=95.82%, *recall*=94.04% dan AUC = 0,979.

Tiara Rahmadani (2021) dalam penelitian terhadap tanyangan televisi berdasarkan opini masyarakat pada media social. Menggunakan metode *K-Nearest Neighbor* Klasifikasi sentimen menghasilkan 3 kategori yaitu sentimen positif, negatif dan netral. Data yang digunakan berupa opini masyarakat terhadap 4 stasiun televisi yaitu Net TV, RCTI, SCTV dan Trans Tv dengan jumlah 2206 data tweet yang dibagi menjadi 70% data latih dan 30% data uji. Hasil diperoleh dengan menghasilkan akurasi terbesar yaitu 72.56% dengan nilai k =3. Dari hasil analisis sentimen dari setiap stasiun televisi didapat respon positif sebesar 69.47% untuk

stasiun Net TV, 72.63% untuk RCTI, 60.53% untuk SCTV, dan 74.32% untuk Trans TV.

Penelitian yang akan dilakukan adalah menggunakan metode *Support Vector Machine* menggunakan fitur kernel linier dan pembobotan TF-IDF terhadap data tweet yang berkaitan dengan pembelajaran tatap muka (PTM) diera pandemi covid-19 untuk mengklasifikasikan sentimen dalam kategori postif dan negatif.

Perbandingan refrensi penelitian diatas dapat dilihat pada Tabel 2.1

Tabel 2.1 perbandingan penelitian sebelumnya

Nama Peneliti	Objek	Metode	Hasil
Risky Maulana (2016)	Opini terhadap tokoh public indonesia	Metode *Support Vector Machine* Dengan LibSVM	presentase analisis sentiment didapat akurasi sebesar 79,5%
Helda ludya (2020)	Opini Publik terhadap Kasus Covid-19	Metode *Support Vector Machine* dan *Radial Basis Function* (RBF).	Persentase Sentimen Positif, Netral, Negatif pada Tweet Opini Covid-19 dengan akurasi 77%
Samsir, dkk (2021)	Pembelajaran daring dimasa pandemi covid-19	Metode *Naïve Bayes*	Klasifikasi pada tweet menunjukkan 30% sentimen positif, 69% sentimen negatif, dan 1% netral.
Rizqi Alfiyati (2021)	Kenaikan iuran BPJS	Metode multiclass *Support Vector Machine*	Dari penelitian tersebut mendapat nilai akurasi 85%.
Primandani Arsi dan Retno Waluyo (2021)	Wacana pemindahan ibu kota Indonesia	Menggunakan metode *Support Vector Machine (SVM)*	Penelitian yang terdiri dari 1.236 *tweets* (404 positif dan 832 negatif) diperoleh akurasi =96,68%, *precision*=95.82%, *recall*=94.04% dan AUC = 0,979
Tiara Rahmadani (2021)	Sentimen terhadap tayangan televisi	Metode K-Neasrest Neighbor	Dari pengujian penelitian menghasilkan akurasi sebesar 72.56% dengan nilai k=3.
Lala Ariska Mulia	Pembelajaran tatap muka (PTM) diera pandemi covid-19	Metode *Support Vector Machine*	Mengetahui Presentase sentimen positif dan negatif dan Menghasilkan nilai akurasi.

2.2 Dasar Teori

2.2.1 Twitter

adalah sebuah situs jejaring sosial yang sedang berkembang pesat saat ini karena pengguna dapat berinteraksi dengan pengguna lainnya dari komputer ataupun perangkat *mobile* mereka dari manapun dan kapanpun. Setelah diluncurkan pada Juli 2006, jumlah pengguna meningkat sangat pesat. Pada September 2010, diperkirakan jumlah pengguna yang terdaftar sekitar 160 juta pengguna (Chiang, 2011).

Pengguna terdapat dari berbagai macam kalangan yang para penggunanya dapat berinteraksi satu sama lain. sebagai sebuah situs jejaring sosial memberikan akses yang memungkinkan pengunanya untuk mengirim, membaca dan membalas pesan berbasis teks hingga 280 karakter yang dikenal dengan sebutan *tweet* atau kicauan. melalui *tweet* pengguna dapat dengan mudah mambagikan tentang apa yang sedang mereka pikirkan, apa yang sedang dilakukan, tentang kejadian yang baru saja terjadi hingga berita terkini serta hal lainnya. Sebagai salah satu layanan jejaring sosial terbesar pengguna telah berkembang pesat serta memiliki berbagai macam manfaat yang dapat diperoleh dari tweet pengguna dimulai dari *even detection* seperti bencana alam, prediksi pemilu, prediksi pergerakan pasar saham, hingga penyebaran wabah penyakit di suatu wilayah serta beberapa bisnis dan organisasi menggunakanya sebagai penyampaian informasi ke *stakeholder.*

2.2.2 Analisis Sentimen

Analisis Sentimen atau *opinion mining* merupakan proses memahami, mengekstrak dan mengolah data tekstual secara otomatis untuk mendapatkan

informasi sentimen yang terkandung dalam suatu kalimat opini. Analisis sentimen dilakukan untuk melihat pendapat atau kecenderungan opini terhadap sebuah masalah atau objek oleh seseorang, apakah cenderung beropini negatif atau positif. (Rozi, Pranomo, & Dahlan, 2012).

Menurut (Liu, 2011) sentiment analysis atau *opinion mining* mengacu pada bidang yang luas dari pengolahan bahasa alami, komputasi *linguistic* dan texts mining yang memiliki tujuan menganalisa pendapat, sentimen, evaluasi, sikap, penilaian dan emosi seseorang apakah pembicara atau penulis berkenan dengan suatu topik, produk, layanan, organisasi, individu, ataupun kegiatan tertentu. Analisis sentimen sangatlah berguna sebagai pemroses penganalisis komentar seperti pendapat kemudian di proses menjadi sesuatu yang lebih bermakna (Palucoumputer, 2015) dari beberapa pendapat diatas dapat disimpulkan bahwa analisis sentimen sebagai pemroses yang dapat mengidentifikasi penilaian dari seseorang mengenai pendapat, sikap opini, serta emosi dari sebuah teks mengenai isu yang sedang terjadi dengan cara mengklasifikasikan dalam kategori kelas positif, negatif maupun netral.

2.2.3 Preprocessing

Pre-processing menjadi tahapan awal dalam melakukan klasifikasi teks untuk mempersiapkan data teks sebelum digunakan pada proses selanjutnya. Tahap *pre-processing* atau praproses data merupakan proses untuk mempersiapkan data mentah sebelum dilakukan proses lain. Pada umumnya, praproses data dilakukan dengan cara mengeliminasi data yang tidak sesuai atau mengubah data menjadi

bentuk yang lebih mudah diproses oleh sistem. Pra-proses sangat penting dalam melakukan analisis sentimen, terutama untuk media sosial yang sebagian besar berisi kata-kata atau kalimat yang tidak formal dan tidak terstruktur serta memiliki noise yang besar. (A Clark, 2003). Proses ini akan mengubah data teks yang sebelumnya data mentah sehingga menghasilkan data informasi teks yang lebih baik. Pada penelitian ini tahap *preprocessing* yang digunakan meliputi:

1. **Case Folding**

   Case folding merupakan proses untuk menyeragamkan karakter pada data. Proses case folding adalah proses mengubah seluruh huruf menjadi huruf kecil. Pada proses ini karakter-karakter 'A'-'Z' yang terdapat pada data diubah kedalam karakter 'a'-'z'.

2. **Text Cleansing**

   Text cleansing merupakan proses pembersihan dokumen untuk mengurangi *noise*. *Text cleansing* sendiri berfungsi untuk menghapus hashtag (#), *username (@username),* HTML, URL,akun,enter,email serta karakter tertentu dan angka yang menggunakan library re dari python.

3. **Tokenizing**

   Tokenizing adalah proses pemecah kalimat menjadi kata yang disebut token untuk kemudian dianalisa.

4. **Stopword Removal**

   Stopword removal merupakan proses menghapus kata yang tidak penting dalam *text.* Hal tersebut dilakukan untuk memperoleh akurasi lebih besar dari pembobotan term.

5. Stemming

Stemming adalah proses untuk mengubah kata yang dihilangkan kata imbuhan awalan dan imbuhan akhiran menjadi kata dasar.

2.2.4 Term Frequency – Invers Document Frequency (TF-IDF)

Tf-Idf adalah perhitungan yang menggambarkan seberapa pentingnya kata (*term*) dalam sebuah dokumen dan korpus. Proses ini digunakan untuk menilai bobot relevansi term dari sebuah dokumen terhadap seluruh dokumen dalam korpus. *Term-frequency* adalah ukuran seringnya kemunculan sebuah term dalam sebuah dokumen dan juga dalam seluruh dokumen di dalam korpus. TF-IDF (*Term Frequency Inverse Document Frequency*) merupakan 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 (Evan, 2014). *Term Frequency* (TF) yaitu semakin tinggi frekuensi kemunculan term pada sebuah dokumen maka akan semakin tinggi juga nilai bobot untuk term itu sendiri. Sementara itu, proses *Inverse Document Frequency* (IDF) merupakan kebalikan dari proses TF. Pada IDF, semakin tinggi frekuensi kemunculan term maka nilai bobot term itu sendiri akan semakin kecil.

1. *Term Frequency* (TF)

   Term Frequency merupakan jumlah kemunculan frekuensi kata pada suatu dokumen (Xia & Chai, 2011). Term Frequency (tf) didefinisikan jumlah kemunculan term t pada dokumen d. pembobotan menggunakan TF dijelaskan pada Persamaan 3.1

$$W_{tf_{t,d}} = \begin{cases} 1 + \log_{10} tf_{t,d}\,, jika\ tf_{t,d} > 0 \\ 0 \qquad\qquad\qquad\quad , jika\ tf_{t,d} = 0 \end{cases}$$

(3.1)

Keterangan:

$W_{tf_{t,d}}$ = Hasil pembobotan $tf_{t,d}$

$tf_{t,d}$ = Banyaknya kemunculan kata t dalam dokumen d

2. *Invers Document Frequency* (IDF)

   Invers Document Frequency merupakan frekuensi kemunculan term pada keseluruhan dokumen teks. *Term* yang jarang muncul pada keseluruhan dokumen teks memiliki nilai *Invers Document Frequency* lebih besar dibandingkan dengan *term* yang sering muncul. Pembobotan menggunakan *Invers Document Frequency* (IDF) dijelaskan pada Persamaan 3.2

$$idf_t = log\left(\frac{N}{df_{(t)}}\right)$$

(3.2)

Keterangan : idf_t = Hasil *inverse* dari df_t

N = Jumlah dokumen teks

$df_{(t)}$ = Jumlah dokumen yang mengandung *term t.*

3. Term Frequency – Invers Document Frequency (TF-IDF)

Nilai tf-idf dari sebuah kata merupakan kombinasi dari nilai tf dan nilai idf dalam perhitungan bobot. Pembobotan TF-IDF dijelaskan pada Persamaan 3.3

$$\text{TF} - IDF_{(t)} = \text{tf} * \text{IDF} \tag{3.3}$$

Keterangan :

$\text{TF} - idf_{(t)}$ = Pembobotan TF_IDF

tf = Hasil Pembobotan $tf_{t,d}$

IDF = *Invers Document Frequency*

2.2.5 Support Vector Machine (SVM)

Support vector machine (SVM) merupakan salah satu metode klasifikasi *supervised learning* yang memprediksi suatu kelas berdasarkan model atau pola dari hasil proses training. *Support vector machine* (SVM) dikembangkan oleh Boser, Guyon dan Vapnik pertama kali diperkenalkan pada tahun 1992 di Annual Workshop on Computation Learning Theory. Konsep dasar metosde SVM sebenarnya merupakan gabungan atau kombinasi dari teori-teori komputasi yang telah ada pada tahun sebelumnya, seperti margin hyperplane (Dyda dan Hart, 1973; Cover, 1965; Vapnik 1964), kernel diperkenalkan oleh Aronszajn pada tahun 1950, Lagrange Multiplier yang ditemukan oleh Joseph Louis Lagrange pada tahun 1766, dan demikian juga dengan konsep-konsep pendukung lainnya.

Menurut Fachrurrazi (2011) (SVM) merupakan suatu teknik untuk melakukan prediksi, baik prediksi dalam kasus regresi maupun klasifikasi. Teknik SVM digunakan untuk mendapatkan fungsi pemisah (hyperplane) yang optimal untuk memisahkan observasi yang memiliki nilai variable target yang berbeeda (William, 2011). Klasifikasi dilakukan dengan mencari hyperpline atau garis pembatas yang memisahkan satu kelas dengan kelas lain. Pada penelitian ini garis tersebut berperan memisahkan tweet yang bersentimen positif (berlabel +1) dengan tweet bersentimen negatif (berlabel -1).

Menurut Nugroho (2003), karakteristik SVM secara umum diarangkum sebagai berikut:

1. Secara prinsip SVM adalah linear classifier.
2. Pattern recognition dilakukan dengan mentransformasikan data pada ruang input *(input space)* ke ruang yang berdimensi lebih tinggi (*feature space*), dan optimisasi dilakukan pada ruang *vector* yang baru tersebut. Hal ini membedakan SVM dari solusi *pattern recognition* pada umumnya, yang melakukan optimisasi parameter pada hasil transformasi yang berdimensi lebih rendah daripada dimensi *input space.*
3. Menerapkan strategi *Structural Risk Minimization* (SRM).
4. Prinsip kerja SVM pada dasarnya hanya mampu menangani klasifikasi dua kelas, namun telah dikembangkan untuk klasifikasi lebih dari dua kelas dengan adanya *pattern recognition.*

Metode *Support Vector Machine* memiliki beberapa keuntungan yaitu:

1. Generalisasi

Generalisasi didefinisikan sebagai kemampuan suatu metode untuk mengklasifikasi suatu pattern atau pola, yang tidak termasuk data yang digunakan dalam fase pembelajaran metode itu.

2. Curse of dimensionality

Curse of dimensionality didefinisikan sebagai masalah yang dihadapi suatu metode *pattern recognition* dalam mengestimasikan parameter dikarenakan jumlah sampel data yang relatif lebih sedikit dibandingkan dengan dimensional ruang vektor tersebut.

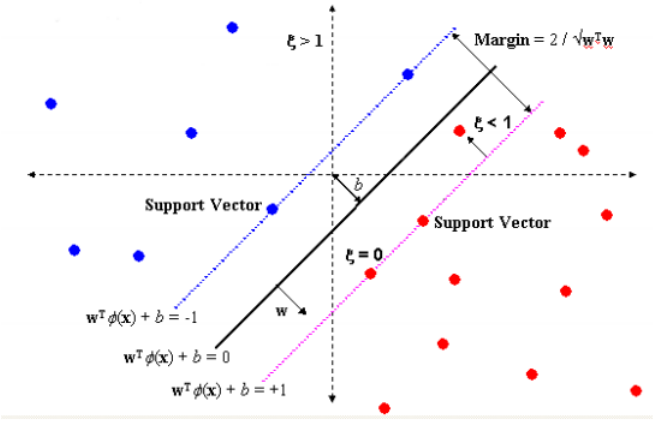
3. Feasibility

SVM dapat diimplementasikan relatif lebih mudah, karena proses penentuan *support vector* dapat dirumuskan dalam *Quadratic Programing* (QP) *problem* (Nugroho, 2003).

Adapun kerugian dari metode *Support Vector Machine* adalah sebagai berikut:

1. Sulit dipakai pada problem berskala besar. Dalam hal ini dimaksudkan dengan jumlah sampel yang diolah.
2. SVM secara teoritik dikembangkan untuk problem klasifikasi dengan dua kelas. Namun dewasa ini SVM telah dimodifikasi

agar dapat menyelesaikan masalah dengan lebih dari dua kelas (Nugroho, 2003).



Sumber: Anandan, Varma,Joy 2014

Gambar 2.1 *Suport Vector Machine*