BAB II
TINJAUAN PUSTAKA DAN LANDASAN TEORI

2.1 Tinjauan Pustaka

Sentiment Analysis merupakan metode analasis yang dipakai untuk mengidentifikasi tentang komentara para pengguna. Menurut Chintala (2012) *sentiment analysis* merupakan metode menganalisis sepotong data untuk emosi manusia. Menurut (GO, Huang, Bhayani, 2009), *sentiment analysis* merupakan area penelitian yang menonjol dan aktif yang didorong dengan pesatnya pertumbuhan media sosial dan kesempatan untuk mengakses opini berharga dari banyak kalangan masalah bisnis, dunia, dan sosial. Menurut Medhat et al (2014, 1093) *sentiment analysis* adalah studi komputasi mengenai pendapat, perilaku, dan emosi seseorang terhadap entitas. Dengan banyaknya pendapat dari pengguna Twitter terhadap suatu masalah, maka penelitian analisis sentimen Twitter sudah banyak dilakukan. Pada kesempatan ini, penulis akan melakukan penjabaran terhadap penelitian yang telah dilakukan.

Pada penelitian yang dilakukan oleh (Elly Indrayuni, 2016) yang berjudul "Analisa Sentimen Review Hotel Menggunakan Algoritma *Support Vector Machine* Berbasis Particle Swarm Optimization" Menurut peneliti hotel merupakan salah satu produk pariwisata yang sangat penting untuk dipertimbangkan baik dari segi fasilitas, pelayanan, ataupun jarak tempuh perjalanan wisata.

Dengan munculnya berbagai macam website wisata yang menyediakan fasilitas bagi pengguna internet untuk menuliskan opini dan pengalaman pribadi secara online. Dengan banyaknya opini yang diberikan oleh pengguna, maka tidak dimungkinkan unutk dibaca secara keseluruhan untuk mengambil suatu keputusan. Maka dari itu dilakukan analisis sentimen.

Dalam penelitian ini, metode yang dipilih adalah Support Vector Machine. Alasan peneliti memilih metode ini karena SVM mampu mengidentifikasi *hyperplane* terpisah yang memaksimalkan margin antardua kelas yang berbeda. Serta keistimewaan SVM yang dapat menerapkan pemisah linier pada input data non linear berdimensi tinggi dengan menggunakan fungsi kernel. Peneliti menggunakan SVM dan Particle Swarm Optimization (PSO) sebagai seleksi fitur untuk meningkatkan nilai akurasi analisa sentimen.

Penelitian ini menggunakan 300 data review hotel yang terdiri dari 150 data opini positif dan 150 data opini negatif berdasarkan review terbaru dari situs www.tripadvisor.com. Dengan menggunakan metode SVM menghasilkan akurasi sebesar 91.33%. Dan penerapan dengan menggunakan seleksi fitur PSO pada algoritma SVM membuat nilai akurasi meningkat menjadi 96.94%.

Pada penelitian yang dilakukan oleh (Jao Allen Banados dan Kurt Junshean Espinosa, 2014) yang berjudul "Optimizing Support Vector Machine in Classifying Sentiments on Product Brands from

Twitter" membahas tentang solusi dalam pengoptimalan SVM dalam mengklasifkasi sentimen terhadap merek produk. Metode klasifikasi yang dipakai adalah *Support Vector Machine* dengan model *unigram* sebagai pilihan fitur. Dalam penelitian ini, data yang akan dikumpulkan adalah sentimen dari pengguna Twitter pada merek tertentu. Data merek ini difokuskan pada merek global terbaik pada tahun 2013.

Sumber data diambil pada Interbrand: Merek Global Terbaik 2013 (http://www.interbrand.com/en/best-global-brands/2013/Best-Global-Brands-2013.aspx). Peneliti mendapatkan 10 merek otomotif teratas, yaitu Toyota, Mercedes-Benz, BMW, Honda, Volkswagen, Ford, Hyundai, Audi, Porsche, dan Nissan. Tweet yang dikumpulkan disaring melalui nama merek dan menggunakan bahasa inggris. Akurasi yang didapatkan dalam menguji data tweet tersebut sebesar 63,5% (2429/3823).

Pada penelitian yang dilakukan oleh (Tiara, Mira Kania Sabariah, Veronikha Effendy, 2015) yang berjudul "Sentiment Analysis in Twitter Using Combination of Lexicon-Based and Support Vector Machine for Assessing the Performance of a Television Program" membahas tentang peningkatan kualitas dari sebuah program televisi. Komentar dari pengguna Twitter dapat melengkapi penilaian program televisi yang biasanya dilakukan dengan menggunakan rating dan terwakilkan dalam bentuk kuantitas. Dengan banyaknya komentar pengguna

Twitter, diharapkan dapat melengkapi asesmen kualitas. Dalam melakukan analisis sentimen, mereka memakai metode SVM yang dikombinasi dengan Lexicon-Based. Kedua metode ini memiliki karakteristik yang berbeda, tetapi dapat saling melengkapi. Metode Lexicon dipakai untuk membuat label tweets yang dijadikan data training di SVM, sehingga tidak akan ada proses pelabelan secara manual.

Penelitian ini menggunakan lima program televisi sebagai data penelitian. Data dipilih secara acak dan dibagi menjadi data *training* dan data *testing*. Total tweets dari lima program ini adalah 2200 tweet yang terbagi atas program A 300 tweets, program B 300 tweets, program C 500, program D 500, dan program E 600. Ada beberapa skenario yang dilakukan dalam pengujian perbandingan, perbandingan kedua data sebagai berikut:

1. Data pelatihan 60%: Data uji 40%
2. Data pelatihan 70%: Data uji 30%
3. Data pelatihan 80%: Data uji 20%
4. Data pelatihan 90%: Data uji 10%

dengan beberapa skenario diatas, nilai rata-rata tertinggi untuk akurasi ada pada perbandingan data pelatihan 90% dan data uji 10%. Setelah semua proses selesai dilakukan, didapatkan program televisi yang didominasi oleh sentimen positif merupkan salah satu yang memiliki nilai keakuratan yang tinggi sebesar 80%.

Pada penelitian yang dilakukan oleh (Huda Al-Shehri, Amani Al-Qarni, Leena Al-Saari, Arwa Batoaq, Haifa Badukhen, Saleh Alrashed, Jamal Alhiyafi, Sunday O. Olatunji, 2017) yang berjudul "Student Performance Prediction Using Support Vector Machine and K-Nearest Neighbor" membahas tentang perkiraan kinerja siswa dalam ujian akhir dengan dua model prediksi. Pada penelitian sebelumnya, dengan kumpulan data yang sama menggunakan algoritma *K-Nearest Neighbor* (KNN) mencapai hasil yang rendah. Sementara algoritma *Support Vector Machine* (SVM) yang merupakan prediksi sangat populer dan kuat teknik jarang digunakan. Dalam penelitian ini, mereka memakai dua model prediksi yaitu KNN dan SVM pada dataset dan membandingkan akurasi dari 2 model tersebut.

Prediksi ini dapat dikelola dengan mencari dari sumber masalah. Masalah tersebut bisa dari kegiata ekstra siswa, masalah keluarga, atau masalah kesehatan. Dataset yang digunakan dalam penelitian dikumpulkan dari dua sekolah menengah Portugis dengan ketentuan mata pelajaran matematika. Data yang digunakan dalam penelitian mereka sebanyak 395 data dari Universitas Minho di Portugal. Penelitian ini mekai dua model algoritma untuk memprediksi final kelas siswa yang jatuh berkisar 0-20.

KNN memiliki tujuh parameter di Weka. Mereka bereksperimen dengan nilai K yang berbeda untuk mencapai koefisien korelasi terbaik. Korelasi tertinggi nilai KNN meningkat menjadi 19 dengan

nilai korelasi sebesar 0,612. 19 KNN mencapai maksimum tinggat akurasinya ketika diterapkan fungsi jarak Manhattan dengan koefisien korelasi 0,67 dan relatif absolut kesalahan 74,73%. Dengan itu untuk mencapai performa terbaik, mereka mengumpulkan parameter optimal yang memberikan akurasi terbaik dengan penurunan kelasahan relatif dan peningkatan nilai korelasi.

Mereka mengoptimalkan parameter SVM dengan menambahkan beberapa parameter. Koefisien dengan parameter default (hanya mengubah jenis SVM ke yipe epsilon-SVR dengan nilai numerik) adalah 0,86 dengan kesalahan absolut 48,89%. Pada saat menguji setiap parameter, perbaikkan terjadi bila memakai kernel liniear dengan mencapai nilai maksimum akurasi koefisien korelasi 0,9 dengan kesalahan absolut 29,5%. Dengan menguji beberapa parameter, mereka memilih tipe parameter kernel linear dengan nilai akurasi tertinggi.

Dengan menggunakan parameter optimal, diperoleh hasil ideal yang dicapai dengan nilai presentase perpecahan 90:10 untuk SVM dan KNN. Hasil akhir menunjukkan SVM sedikit melebihi KNN untuk masalah ini engan nilai koefisien korelasi 19,92%. Berdasarkan hasil yang mereka dapatkan, SVM maupun KNN akan sesuai dengan masalah ini.

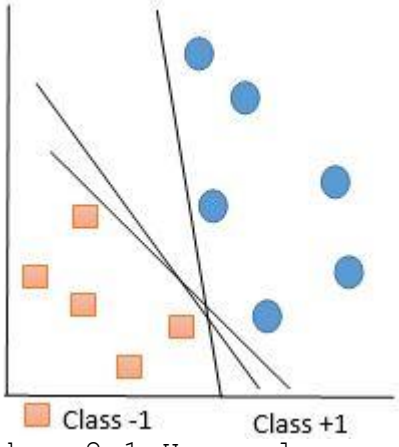
2.2 Landasan Teori

2.2.1 Metode Support Vector Machine

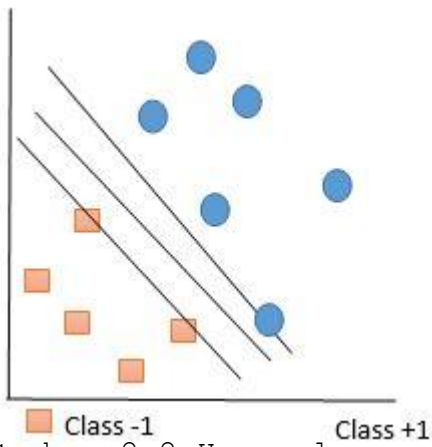
Support Vector Machine pertama kali diperkenalkan oleh Vapnik pada tahun 1992 sebagai rangakaian harmonis konsep – konsep unggulan dalam

bidang pattern recognition. Konsep dasar SVM merupakan kombinasi dari teori komputasi yang telah ada puluhan tahun sebelumnya, yaitu margin hyperplane (Duda & Hart tahun 1973, Cover tahun 1965, Vapnik 1964, dsb). *Support Vector Machine* merupakan salah satu metode yang terbimbing. Metode terbimbing yang dimaksud adalah metode yang membutuhkan data *training* dan data *testing* dalam uji coba. Menurut Basari et al (2013, 453) Support Vector Machine adalah metode untuk menganalisa data dan mengenali pola yang bisa digunakan untuk pengklasifikasian. Dalam pemakaian metode ini, harus memakai pelabelan. Pelabelan tersebut adalah kalimat positif atau kelas negatif.

Tujuan dari metode ini adalah menemukan *hyperplane* yang optimal yang memiliki margin maksimal. Margin tersebut adalah jarak antara *hyperplane* dengan titik terdekat setiap kelas. Dewasa ini, SVM telah digunakan dalam masalah dunia nyata dan memberikan solusi yang lebik baik secara keseluruhan.



Gambar 2.1 Hyperplane SVM(1)



Gambar 2.2 Hyperplane SVM(2)

Pada gambar memperlihatkan beberapa pattern yang merupakan anggota dari dua buah class: +1 dan –1. Pattern yang tergabung pada class –1 disimbolkan dengan warna merah (kotak), sedangkan pattern pada class +1, disimbolkan dengan warna biru(lingkaran). Problem klasifikasi dapat diterjemahkan dengan usaha menemukan garis (*hyperplane*) yang memisahkan antara kedua kelompok tersebut. Berbagai alternatif garis pemisah (*discrimination boundaries*) ditunjukkan pada gambar 2.1. *Hyperplane* pemisah terbaik antara kedua kelas dapat ditemukan dengan mengukur margin *hyperplane* tersebut. dan mencari titik maksimalnya. Margin adalah jarak antara *hyperplane* tersebut dengan pattern terdekat dari masing-masing class. Pattern yang paling dekat ini disebut sebagai *support vector*. Garis solid pada gambar 2.2 menunjukkan hyperplane yang terbaik, yaitu yang terletak tepat pada tengah-tengah kedua class, sedangkan titik merah dan kuning yang berada dalam lingkaran hitam adalah support vector. Usaha untuk mencari lokasi hyperplane ini merupakan inti dari proses pembelajaran pada SVM. Diasumsikan kedua kelas -1 +1 dapat terpisah secara sempurna oleh hyperplane berdimensi d, yang didefinisikan

$$w.x + b = 0$$

Pattern yang termasuk class -1(negatif) dapat didefinisikan dengan persamaan sebagai berikut

$$w.x_i + b \leq -1$$

Sedangkan pattern yang termasuk class +1(positif) didefinisikan dengan persamaan sebagai berikut

$$w.x_i + b \geq +1$$

2.2.2 Pemenuhan Kebutuhan Dasar

Pemenuhan kebutuhan dasar merupakan salah satu penyelenggaraan penanggulangan bencana pada tahap tanggap darurat, meliputi penyediaan:

a. Kebutuhan air bersih dan sanitasi
b. Pangan
c. Sandang
d. Pelayanan kesehatan
e. Pelayanan psikososial
f. Penampungan dan tempat hunian

2.2.3 Kernel SVM

Pada umumnya masalah di dunia jarang yang bersifat linear, kebanyakan bersifat non liniear. Agar dapat menyelesaikan masalah non linear, SVM menggunakan fungsi kernel. Fungsi kernel ini memetakan class yang ada pada input space berdimensi dua ke dalam vektor baru yang berdimensi tinggi.

Tabel 2.1 Kernel dalam SVM

Jenis Kernel	Definisi
Liner	$K_{(x,y)} = x.y$
Polynomial	$K_{(x,y)} = (x.y + c)^d$
Radial Basis Function RBF)	$K_{(x,y)} = exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right)$
Sigmoid	$K_{(x,y)} = \tanh((k\langle x.y\rangle + \vartheta)$

2.2.4 RStudio

RStudio merupakan open-source pengembangan integrasi dan gratis untuk R, Bahasa pemrograman untuk komputasi statistic dan grafik. RStudio didirikan oleh JJ Allaire pencipta Bahasa pemrograman ColdFusion. Hadley Wickham adalah kepala ilmuwan di RStudio. RStudio tersedia dalam 2 platform, RStudio Desktop dan RStudio Server. RStudio Desktop merupakan program yang dijalankan secara lokal. RStudio Server yang memungkinkan mengakses RStudio menggunakan browser web.

2.2.5 K-fold Cross Validation

K-fold Cross Validation merupakan teknik yang memecah dataset sebanyak k subset. Satu dari subset akan dijadikan sebagai data testing dan k-1 lainya menjadi data training. Itu terjadi berulang-ulang sebanyak K kali. k-fold cross validation ini memakan proses komputasi yang lebih besar karena harus melakukan k kali proses. Namun keuntungannya mendapatkan nilai rata-rata tertinggi.

2.2.6 Data Mining

Menurut (Davies, 2004) *Data Mining* adalah penambangan atau penemuan informasi baru dengan mecari pola atau aturan tertentu dari sejumlah data yang sangat besar. Menurut (Santoso, 2007) KDD adalah kegiatan yang meliputi pengumpulan, pemakaian data, historis untuk menemukan keteraturan, pola atau hubungan dalam set data berukuran besar (Santoso, 2007). Ada tahapan yang terjadi pada data mining (Fayyad, 1996), yaitu:

1. Data Selection

   Pada tahap ini, terjadi pemilihan data dari sekumpulan data.

2. Preprocessing / Cleaning

   Pada tahap ini, terjadi proses membuang dupikasi data, memeriksa data yang inkonsisten dan memperbaiki kesalahan data.

3. Transformation

Coding merupakan proses transformasi pada data yang telah dipilih, sehingga data tersebut sesuai untuk proses data mining.

4. Data Mining

Pada tahap ini, terjadi pencarian pola dengan menggunakan metode tertentu.

5. Interpretation / Evaluation

Pada tahap ini informasi yang dihasilkan perlu ditampilkan dalam bentuk yang mudah dimengerti oleh pihak yang berkepentingan