BAB II

TINJAUAN PUSTAKA DAN DASAR TEORI

1.1.Tinjauan Pustaka

Imelda (2015) dalam penelitian tentang penerapan metode *Support Vector Machine* pada klasifikasi *tweet* dengan menggunakan *Kernel Radial Basis Function (RBF)* agar *tweet* yang ada tidak bercampur antara iklan dan tidak iklan. Parameter yang digunakan berupa nilai C dan γ yang kemudian dihasilkan nilai akurasi tertinggi yaitu 99,12% sehingga dapat diterapkan agar memberikan bantuan kepada pengguna dalam mengelola *tweet*, terutama *tweet* iklan.

Umi Rofiqoh (2017) dalam penelititan tentang penerapan metode *Support Vector Machine* pada klasifikasi *tweet* tingkat kepuasan pengguna penyedia layanan telekomunikasi seluler dengan menggunakan nilai degree sebesar 2 dan nilai konstanta *learning rate* 0,0001. Dari penelitian tersebut dihasilkan akurasi sistem sebesar 79% sehingga memudahkan untuk mengetahui tingkat kepuasan pengguna penyedia layanan telekomunikasi seluler dengan hasil klasifikasi berupa sentimen positif atau negatif.

Al Hafiz (2018) dalam penelitian tentang klasifikasi *tweet e-commerce* menggunakan metode *Support Vector Machine* dengan fitur *kernel Radial Basis Function* (RBF). Dari penelitian ini dihasilkan klasifikasi berupa tweet terkait transaksi *e-commerce* dan non transaksi *e-commerce* dan diketahui parameter terbaik adalah pasangan parameter C=0,9 dan γ=0,8 dengan nilai akurasi sebesar 96,1%

Arsya Monica Pravina (2019) melakukan penelitian dengan topik Analisis Sentimen pada Dokumen Twitter, dimana objek dari penelitian ini adalah opini maskapai penerbangan pada dokumen twitter menggunakan metode *Support Vector Machine* serta klasifikasi sentimen dengan fitur *Lexicon Based* yang dapat menerima opini berbahasa lain selain Bahasa Indonesia.

Dari penelitian tersebut dihasilkan klasifikasi opini positif, negatif dan netral dengan menggunakan parameter C dan γ dan memberikan hasil akurasi sebesar 40%

Rian (2020) pada penelitiannya terhadap layanan indihome berdasarkan twitter menggunakan metode *Support Vector Machine* , diperoleh nilai positif sebesar 18,4% dan hasil negatif sebesar 81,6% serta akurasi sebesar 87% dengan ketepatan antara hasil prediksi dengan data sebenarnya (precision) sebesar 86%, tingkat keberhasilan sistem dalam memprediksi sebuah data (recall) sebesar 95%, tingkat kesalahan semua data yang diprediksi (error rate) sebesar 13%, sedangkan untuk nilai perbandingan rata-rata precision dan recall (f1- score) adalah sebesar 90%.

Penelitian yang akan dilakukan adalah menggunakan metode *Support Vector Machine* menggunakan fitur *Kernel Radial Basis Function* terhadap data tweet terkait dengan sistem zonasi untuk mengklasifikasikan sentimen kedalam kategori positif, negatif dan netral.

Tabel 2.1 berisikan referensi penelitian yang menggunakan metode *Support Vector Machine* dengan fitur tambahan untuk mengklasifikasikan sentimen pada data *tweet.*

Tabel 2. 1 Perbandingan Penelitian

No	**Nama Peneliti**	**Metode dan teknologi**	**Objek**	**Hasil**
1	Imelda A.Muis dan Muhammad Affandes, M.T (2015)	*Support Vector Machine* dengan *Radial Basis Function*	*Tweet* iklan pada tweeter	Klasifikasi *tweet* iklan dan tidak iklan dengan nilai akurasi 99.12%
2	Umi Rofiqoh, dkk (2017)	*Support Vector Machine*	Opini Maskapai	Klasifikasi *tweet* positif dan *tweet* negatif dengan

		dengan *Lexicon Based Features*	Penerbangan pada Dokumen Twitter	akurasi sebesar 79%
3	Al Hafiz, dkk (2018)	*Support Vector Machine* dengan *Radial Basis Function*	*Tweet e-commerce*	klasifikasi *tweet* transaksi *e-commerce* dan non transaksi *e-commerce* dengan akurasi sebesar 96.1%
4	Arsya Monica Pravina, dkk	*Support Vector Machine* dengan *Lexicon Based Features*	Opini Maskapai Penerbangan pada Dokumen Twitter	klasifikasi opini positif, negatif dan netral dengan akurasi sebesar 40%
5	Rian Tineges, dkk (2020)	*Support Vector Machine*	Sentimen layanan Indihome	Presentase sentimen positif dan negatif dengan nilai akurasi sebesar 87%

1.2.Dasar Teori

1.2.1. Twitter

Twitter merupakan media sosial yang memungkinkan pengguna untuk mengekspresikan opini dan perasaan mereka mengenai banyak isu atau permasalahan (Hamdan, Bellot & Bechet, 2015). Berbeda dengan media sosial yang lain yang harus menjadi teman terlebih dahulu baru dapat berinteraksi, Twitter memungkinkan antar pengguna tetap terhubung walaupun mereka tidak saling berteman (Windasari, Uzzi & Satoto, 2017).

1.2.2. Analisis Sentimen

Analisis sentimen merupakan salah satu cabang ilmu dari text mining, natural language program, dan artificial intelegence. Proses yang dilakukan oleh analisis sentimen untuk memahami, mengekstrak, dan mengolah data teks secara otomatis sehingga menjadi suatu informasi yang bermanfaat (Akbari, et al., 2012).

Key idea dari SVM adalah untuk menemukan permukaan keputusan (Hyperlane) yang maksimal dari setiap titik data . Untuk melakukan training mesin yang didukung oleh vector atau biasa disebut Support Vector Machine (SVM) memerlukan solusi Quadratic Programming (QP) yang sangat besar. Quadratic Programming adalah masalah matematika untuk menemukan vector “x” yang meminimalkan fungsi kuadrat , dengan melakukan pembagian kelas menggunakan hyperplane maka masing-masing kelas positif, netral dan negatif dapat dibagi berdasarkan area masing-masing sehingga ketika terdapat data baru dapat ditentukan kelasnya berdasarkan area positif, netral maupun negatif (Nurirwan, 2015).

1.2.3. *Preprocessing*

Text preprocessing merupakan tahapan sangat penting dalam melakukan proses klasifikasi data teks. Tujuan dilakukannya text preprocessing yaitu untuk menghilangkan noise, menyeragamkan bentuk kata dan mengurangi volume kata. Berikut tahapan didalam preprocessing data teks.

1. Case Folding

Langkah case folding digunakan untuk menyeragamkan semua huruf dalam teks menjadi huruf kecil atau huruf besar. Strategi umum yang digunakan adalah mengubah menjadi huruf kecil.

2. Cleaning

Langkah cleaning digunakan untuk membersihkan noise dari teks. Dalam penelitian ini, pembersihan dilakukan dengan mengubah noise menjadi karakter spasi. Entitas tweet yang dibersihkan meliputi URL, mention dan hashtag yang dikenali dengan ciri berikut. a. URL diawali ‘http’, ’https’, ’ftp’ atau ’file’. b. Mention diawali simbol ‘@’ yang menunjukkan sebuah akun Twitter. c. Hashtag diawali karakter ‘#’. Sementara itu, karakter yang dibersihkan meliputi karakter HTML, emoticon, angka dan tanda baca.

3. Tokenizing

Langkah tokenisasi digunakan untukmemisahkan teks menjadi token-token yang bermakna. Dalam penelitian ini, teks dipisahkan menjadi kata berdasarkan karakter pemisahnya yaitu spasi.

4. Stopword

Stopword merupakan daftar kata umum yang tidak memiliki arti penting dan tidak digunakan. Pada proses ini kata umum akan dihapus untuk mengurangi jumlah kata yang disimpan oleh sistem (Manning, et al., 2009).

5. Stemming

Langkah stemming digunakan untuk mengubah kata ke dalam bentuk dasarnya (root). Dalam penelitian ini, digunakan lib sastrawi pada phyton

6. Ekstraksi Fitur

Fitur (ciri) yang digunakan dalam klasifikasi ini adalah kata, dengan jenis fitur unigram. Tidak semua kata hasil preprocessing yang akan digunakan. Oleh sebab itu, dilakukan penilaian pentingnya setiap kata dengan menggunakan Document Frequency (DF).

1.2.4. *Term Frequency-Inverse Document Frequency*

Term Frequency-Inverse Document Frequency (TF-IDF) adalah metode yang digunakan untuk menghitung bobotan setiap kata yang telah diekstrak. Penggunaan metode ini umumnya dilakukan untuk mengihitung kata umum yang ada pada information retrieval. Model pembobotan TF-IDF merupakan metode yang mengintegrasikan model *term frequency* (tf) dan *inverse document frequency* (idf). *Term frequency* (tf) merupakan proses untuk menghitung jumlah kemunculan term dalam satu dokumen dan *inverse document frequency* (idf) digunakan untuk menghitung *term* yang muncul di berbagai dokumen (komentar) yang dianggap sebagai *term* umum, yang dinilai tidak penting (Akbari, et al., 2012).

Tahapan pembobotan dengan TF-IDF adalah:

1. *Term Frequency (tf)*

Term frequency atau tf merupakan jumlah kemunculan atau frekuensi kata pada suatu dokumen. Sementara W_{tf} adalah jumlah bobot dari tf yang telah dihitung dengan logaritma.

Perhitungan *Term Frequency* dilakukan dengan persamaan 2.1:

$$Wtf_{t,d} = \{ 0, {}^{1 + \log_{10} tft,d \text{ , if } tft,d > 0} \quad \text{.....(2.1)}$$

2. *Document Frequency* (df)

Document Frequency (df) merupakan frekuensi atau jumlah dokumen yang mengandung suatu kata.

3. *Inverse Document Frequency* (idf)

Inverse Document Frequency (idf) adalah bobot kebalikan dari bobot document frequency. Kata yang jarang muncul di banyak dokumen mempunyai bobot Inverse Document Frequency yang tinggi.

Perhitungan dari *Inverse Document Frequency* (idf) dilakukan dengan persamaan 2.2:

$$idf_t = log_{10}(N/df_t)\ldots\ldots(2.2)$$

Keterangan:

N : jumlah dokumen teks.

dft : jumlah dokumen yang mengandung suatu kata t.

4. *Term Frequency-Inverse Document Frequency* (tf-idf)

   Pembobotan ini adalah hasil perkalian dari pembobotan *term frequency* dan *inverse document frequency* dari suatu *term*.

   Perhitungannya dapat dilihat melalui persamaan 2.3:

$$W_{t,d} = Wtf_{t,d} \times idf_t \ldots..(2.3)$$

   Keterangan :

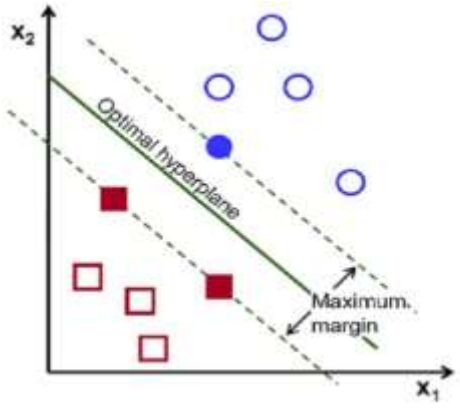
   $wtft,d$: Term Frequency.

   $idft$: Inverse Document Frequency.

1.2.5. *Support Vector Machine*

Support Vector Machine (SVM) adalah suatu teknik yang relatif baru untuk melakukan prediksi, baik dalam kasus klasifikasi maupun regresi. *Support Vector Machine* masuk kelas *supervised learning*, dimana dalam implementasinya perlu adanya tahap pelatihan menggunakan *sequential training* SVM dan disusul tahap pengujian (Santosa, 2015). Konsep klasifikasi dengan Support Vector Machine adalah mencari *hyperplane* terbaik yang berfungsi sebagai pemisah dua kelas data. *Support Vector Machine* mampu

bekerja pada dataset yang berdimensi tinggi dengan menggunakan kernel trik. *Support Vector Machine* hanya menggunakan beberapa titik data terpilih yang berkontribusi (support vector) untuk membentuk model yang akan digunakan dalam proses klasifikasi. Ilustrasi metode *Support Vector Machine* ditunjukkan pada Gambar 2.1.



Gambar 2. 1 Ilustrasi metode *Support Vector Machine*

Metode *Support Vector Machine* didefinisikan pada persamaan 2.4 dan 2.5:

$$f(x) = w.\ x + b \text{.....(2.4)} \quad \text{atau}$$

$$f(x) = \sum aiyiK(x, xi\ ) + b\ m\ i{=}1 \text{ .....(2.5)}$$

Keterangan :

w : parameter *hyperplane* yang dicari (garis yang tegak lurus antara garis *hyperplane* dan titik *support vector*)

x : titik data masukan *Support Vector Machine*

ai : nilai bobot setiap titik data

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

b : parameter *hyperplane* yang dicari (nilai bias)

1.2.6. *Pandas Library*

Pandas merupakan sebuah *open source python package/library* dengan lisensi BSD yang menyediakan banyak perkakas untuk kebutuhan data analisis, manipulasi dan pembersihan data. Pandas mendukung pembacaan dan penulisan data dengan media berupa *excel spreadsheet*, CSV, dan SQL yang kemudian akan dijadikan sebagai objek python dengan *rows* dan *columns* yang disebut *data frame* seperti halnya pada tabel statistik.

1.2.7. *Scikit-learn*

Scikit – Learn adalah modul python yang mengintegrasikan berbagai algoritma pembelajaran mesin *state-of-the-art* untuk masalah yang diawasi dan tidak diawasi skala menengah. Paket ini berfokus pada membawa pembelajaran mesin ke non-spesialis menggunakan bahasa tingkat tinggi tujuan umum. Penekanan diberikan pada kemudahan penggunaan, kinerja, dokumentasi, dan konsistensi API. Ini memiliki ketergantungan minimal dan didistribusikan dibawah lisensi BSD yang disederhanakan, mendorong penggunaannya baik dalam aturan akademis dan komesial. (Pedregosa Fabiann, 2011).

Library dibangun diatas *SciPy* (*Scientific Python*) yang harus diinstal sebelum menggunakan scikit – learn. Tumpukan ini meliputi :

1. Numpy : Paket array n – dimensi dasar
2. Scipy : Pustaka dasar untuk komputasi ilmiah
3. Matplotlib : Komprehensif 2D / 3D
4. Ipyton : Peningkatan konsol interaktif
5. Sympy : Matematika simbolik
6. Pandas : Struktur dan analisis data.

Ekstensi atau modul untuk *Scipy care* secara konvensional diberinama *Scikits*. Modul ini menyediakan algoritma pembelajaran dan diberi nama *scikit – learn* (Brownlee Jason, 2014).