BAB III
METODOLOGI PENELITIAN

3.1 Subyek dan Obyek Penelitian

Subyek dalam penelitian ini adalah persepsi pengguna jasa kurir SiCepat terhadap pelayanan SiCepat melalui review di Twitter. Obyek penelitian ini adalah salah satu jasa kurir yang ada di Indonesia yaitu SiCepat Ekspress.

3.2 Alat dan Bahan Penelitian

3.2.1 Alat Penelitian

Pada penelitian ini menggunakan alat penelitian berupa perangkat keras dan perangkat lunak sebagai tools untuk merancang aplikasi tersebut, yaitu:

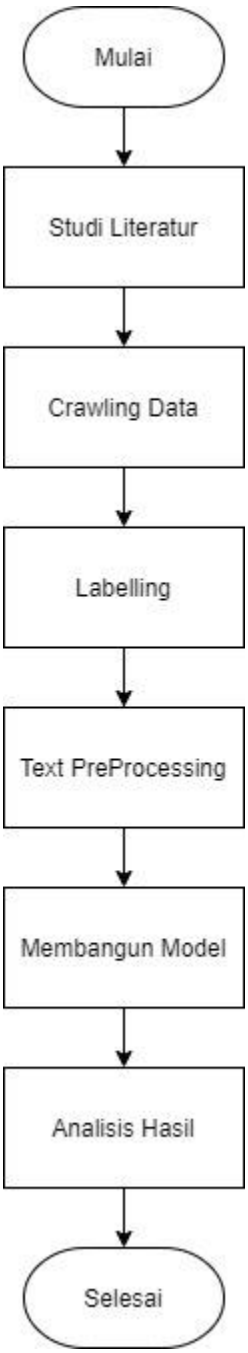
1. Perangkat Keras
   a. Processor Intel(R) Core (TM) i3
   b. RAM 8 GB.
   c. Mouse dan Keyboard
2. Perangkat Lunak
   a. Windows 10 Enterprise
   b. Microsoft Office 2019
   c. Jupyter Notebook
   d. Python 3.8.3

3.2.2. Bahan Penelitian

Bahan penelitian yang akan digunakan yaitu Data dari akun twitter @Sicepat_ekspres yang diambil dengan metode twitter scrapper.

3.3 Diagram Alir Penelitian

Diagram alir/ *flowchart* penelitian ditunjukkan dalam Gambar 3.1, sebagai berikut:



Gambar 3.1 Diagram Alir Penelitian

3.3.1 Studi Literatur

Studi literatur merupakan tahap yang akan dilakukan oleh peneliti untuk mencari berbagai informasi dari buku, jurnal, paper yang berhubungan dengan Sentimen Analisis. Pada tahap ini peneliti mencari sebuah metode yang akan digunakan untuk penelitian ini, dengan melihat dari penelitian-penelitian yang sudah dilakukan oleh orang lain.

3.3.2 Crawling Data

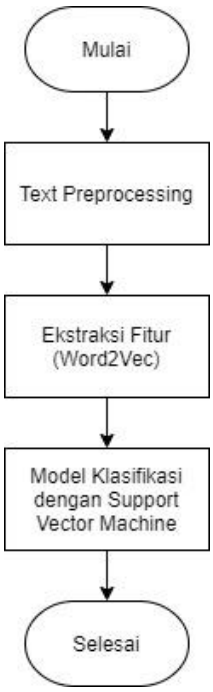
Pada tahap ini diperlukan pengumpulan data dari web twitter.com dengan keyword @SiCepat_ekspres, Dataset diambil dari media sosial Twitter menggunakan API Twitter dan library tweepy dari Python data yang diambil sejumlah 945 data.

3.3.3 Labelling

Labelling adalah klasifikasi dari tweet, apakah tweet tersebut memiliki sentimen bernilai positif, negatif, dan netral[21]. Proses labelling ini dilakukan secara manual, proses ini dilakukan oleh satu orang. Tweet yang dilabeli merupakan tweet asli yang belum diolah.

3.3.4 Membangun Model

Model klasifikasi sentimen menggunakan SVM dan ektraksi fitur dengan Word2Vec ditunjukakn oleh Gambar 3.2



Gambar 3.2 Alir Diagram Membangun Model

1. Text Preprocessing

   Text Preprocessing berguna untuk menyeleksi data dan mengubahnya menjadi data yang terstruktur dan merupakan tahapan awal untuk mengubah struktur isi dari data untuk menjadi format yang sesuai agar dapat diproses oleh Word2Vec. Adapun beberapa tahap *text preprocessing* yang dilakukan pada penelitian kali ini sebagai berikut:

   a. Cleaning yaitu membersihkan noise dari teks, Pembersihan dilakukan dengan mengubah noise menjadi karakter spasi. Entitas tweet yang dibersihkan meliputi URL, mention dan hashtag.
   b. Filtering adalah tahap menghapus kata-kata yang kurang bermakna atau tidak memiliki arti seperti kata: saya, dan, atau.
   c. Tokenizing untuk memisahkan teks menjadi token atau kata yang bermakna.
   d. Stemming untuk mengubah kata ke dalam bentuk dasar (root) pada teks[22].

2. Vektorisasi Menggunakan Word2vec

Analisis sentimen terdapat ekstraksi fitur. Ekstraksi fitur yaitu tahap di mana kata, dapat menjelaskan sentimen pada dataset diekstraksi menjadi fitur atau aspek[6]. Word2vec digunakan untuk ekstraksi fitur. Word2vec merupakan algoritma word embedding, yaitu pemetaan dari kata menjadi vector. Vector ini nantinya digunakan untuk berbagai macam tugas Natural Language Processing. Pada Word2vec pada pembobotan kata menggunakan nilai rata-rata vector yang mewakili kata tersebut. Pada tahap ini akan dilatih data tweet @sicepat_ekspres berbahasa Indonesia dengan metode word2vec. Pelatihan word2vec memerlukan data dengan jumlah kata yang lengkap, sehingga penelitian ini menggunakan data tweet @sicepat_ekspres.

Word2vec digunakan untuk merepresentasikan vektor kata serta memiliki dua arsitektur pemodelan, yaitu continous bag-of-word (CBOW) dan Skip-gram. Pada penelitian kali ini penulis menggunakan arsitektur Skip-gram. Arsitektur Skip-gram dapat mengenali tiga layer, layer yang pertama yaitu layer input, layer yang kedua yaitu hidden layer, dan layer yang terakhir yaitu output layer. Pada input layer di hidden layer terdapat matriks weight (W) berasal dari nilai random, layer ini digunakan untuk mengaktifkan hidden layer. Perhitungan hidden layer dengan matriks weight'(W') dihasilkan dari Output layer.

Windows Size	Text	Skip-grams
2	Kecewa dengan pelayanan sicepat_ekspres input target	Kecewa, dengan Kecewa, pelayanan
	Kecewa dengan pelayanan sicepat_ekspres target input target	dengan, kecewa dengan, pelayanan dengan, sicepat_ekspres
	Kecewa dengan pelayanan sicepat_ekspres target input target	pelayanan, sicepat_ekspres pelayanan, dengan pelayanan, kecewa
	Kecewa dengan pelayanan sicepat_ekspres target input	sicepat_ekspres, pelayanan sicepat_ekspres, dengan

Gambar 3.3 Contoh Skip-gram tweet Sicepat

Vektor kalimat dari data tweet sicepat: Kecewa dengan pelayanan sicepat, maka hasil dari kerja skip-grams seperti pada gambar 3.4

Cara kerja ekstraksi fitur word2vec seperti berikut :

Pertama Kata-kata yang bersumber dari teks diubah ke vektor menggunakan one-hot encoding, didapat vektor dan dimensi

Tabel 3.1 Hasil Encoding Setiap Term dalam Opini

Id_term	term_Dengan	term_Kecewa	term_Pelayanan	term_Sicepat
0	0	1	0	0
1	1	0	0	0
2	0	0	1	0
3	0	0	0	1

Dari Tabel 1, encoder merupakan hasil dari vektor kata. Contoh, kata dengan menjadi vector [0 1 0 0]. Urutan term pada hasil encoding tidak selalu urut seperti contoh vektor kalimat yang ada diatas. Selanjutnya, vektor w(t) hasil dari encoding tersebut menjadi vektor input dan target output. W (t) adalah input dan menjadi output.

Kedua: Vektor input dari term akan menjadi hidden layer dari jumlah |v| yang artinya jumlah fitur yang digunakan. Hidden layer ini merupakan perkalian antara bobot vektor W[|v|, N] dan input vector w(t).

Ketiga: *Hidden layer* tidak ada fungsi aktivasi, karena H[1,k] akan langsung dilewatkan ke output layer.

Keempat: *Output* layer akan dihitung dalam perkalian antara H[1, N] dan W'[N, |v|] maka hasilnya vektor U.

Kelima: Menghitung probabilitas pada vektor dengan fungsi *softmax* dengan persamaan.

$$p(w_{c,j} = wO_{,c}|wI) = \frac{\exp u_{cj}}{\sum_{j=1}^{V} \exp u_j} \tag{3.1}$$

Keterangan:

$w(c,j)$: kata ke-j yang diprediksi pada posisi context ke-c;

$w(O,c)$: kata aktual yang ada pada posisi context ke-c;

$w(I)$: kata input;

$u\ (c,j)$: nilai ke-j pada vektor U untuk memprediksi kata pada posisi konteks ke-c[11].

Tabel 3.2 Word Output dan Vektor

Kata	V0	V1	V2	V3
Kecewa	0.07392003	-0.09522382	0.07700656	-0.0732253
Dengan	-6.5396048e-02	-2.6099820e-05	1.0330487e-02	4.5145866e-02
Pelayanan	0.06874003	-0.07227158	-0.05891787	0.03351511
Sicepat	-0.12403438	0.09518109	0.11903343	0.01476164

Pada Tabel 2, kolom v0, v1,v2,v3 merupakan fitur yang terbentuk sesuai ukuran fitur. Matrik vektor hasil word2vec merepresentasikan seluruh vektor term pada korpus yang digunakan.

Keenam: Pada Langkah terakhir memilih vektor output untuk model klasifikasi. Metode yang digunakan yaitu *average base*, mencari nilai rata-rata dari vektor-vektor sebagai kata penyusun opini untuk memprediksi jenis sentimennya.

Misalnya:

Opini a: Kecewa dengan sicepat

Opini b: Kecewa pelayanan sicepat

Pada Tabel 2, nilai average untuk opini a pada Tabel 3 dan opini b pada Tabel 4.

Nilai rata-rata tiap fitur akan digunakan sebagai atribut dalam membangun model klasifikasi.

Tabel 3.3 Word Output dan Input Opini a

Kata	V0	V1	V2	V3
Kecewa	0.07392003	-0.09522382	0.07700656	-0.0732253
Dengan	-6.5396048e-02	-2.6099820e-05	1.0330487e-02	4.5145866e-02
Sicepat	-0.12403438	0.09518109	0.11903343	0.01476164
Rata-Rata	-0.038503466	-2.29433E-05	0.068790159	-0.004439265

Tabel 3.4 Word Output dan Input Opini b

Kata	V0	V1	V2	V3
Kecewa	0.07392003	-0.09522382	0.07700656	-0.0732253
Pelayanan	0.06874003	-0.07227158	-0.05891787	0.03351511
Sicepat	-0.12403438	0.09518109	0.11903343	0.01476164
Rata-Rata	0.00620856	-0.02410477	0.045707373	-0.008316183

3. Model Klasifikasi dengan Support Vector Machine

   Alir diagram model klasifikasi SVM ditunjukkan oleh Gambar 3.4

Gambar 3.4 Alir Diagram *Support Vector Machine*

Proses klasifikasi dilakukan untuk menentukan sentimen sebuah tweet opini. Metode dalam penelitian ini menggunakan *Support Vector Machine* karena termasuk klasifikasi yang dapat mengubah dokumen teks menjadi vector sebelum diklasifikasi. Dataset diperoleh dari proses sebelumnya, maka akan dirubah ke bentuk matriks vektor untuk penelitian selanjutnya.

1. Nilai X adalah fitur-fitur setiap data hasil word2vec, sedangkan Y-nya adalah kelas/ label setiap komentarnya. Memasukan Nilai x training vector dan nilai y train digunakan sebagai target kelas.
2. Melakukan perhitungan menggunakan kernel linear yang disediakan di SVM.
3. Optimasi dengan nilai parameter C untuk keperluan misclassification. Untuk nilai C yang besar, optimasi akan memilih hyperplane yang lebih kecil.
4. Hasil akurasi untuk melakukan perhitungan dalam proses klasifikasi data, menggunakan metode *Confusion matrix* untuk menentukan nilai akurasi, presisi dan recall.

Langkah-langkah metode *Support Vector Machine* mengacu pada penelitian Tineges, Triayudi, dan Sholihati [23]

a. Mencari kata yang sering muncul pada tweet.
b. Mencari inisialisasi nilai α= 0.5, C=1, λ=0.5, gamma=0.5 dan epsilon=0.001.
c. Menentukan nilai matriks pada Persamaan:

$$D_{ij} = y_i y_j \left(K^{(\vec{x_i} \cdot \vec{x_j})} + \lambda^2\right) \tag{3.2}$$

Keterangan :

D_{ij} : elemen matriks data ke-ij

y_i : kelas atau label data ke-i

y_j : kelas atau label data ke-j

λ : turunan batas teoritis

$K(\vec{x_i} \cdot \vec{x_j})$: fungsi kernel

Tabel 3.5 Perhitungan SVM Linier

x_1	x_2	Kelas (y)	Support Vector (SV)
1	1	1	1
1	-1	-1	1
-1	1	-1	1
-1	-1	-1	0

Pada Tabel 3.5 terdapat empat data masing masing memiliki nilai x_1, x_2 kelas (y) dan Support Vector (SV). Terdapat dua fitur yaitu x_1 dan x_2 maka w juga memiliki 2 fitur yaitu w_1 dan w_2 . Disini peneliti akan menentukan Hyperplane dengan Rumus yang digunakan sebagai berikut :

Meminimalkan nilai

$$\frac{1}{2}||w||^2 = \frac{1}{2}(w_1^2 + w_2^2) \tag{3.3}$$

Syarat

$y_1(w.x_1 + b) \geq 1, \quad i = 1,2,3, \dots \dots N$

$y_1(w_1.x_1 + w_2.x_2 + b) \geq 1$

Contoh perhitungan

$y_1(w_1.x_1 + w_2.x_2 + b) \geq 1$

$1(w_1.1 + b) \geq 1$

$(w_2.x_2 + b) \geq 1$

Hasil

1. $(w_1 + w_2 + b) \geq 1,\ untuk\ y_1 = 1, x_1 = 1,\ x_2 = 1$
2. $(-w_1 + w_2 - b) \geq 1,\ untuk\ y_2 = -1, x_1 = 1,\ x_2 = -1$
3. $(w_1 - w_2 - b) \geq 1,\ untuk\ y_3 = -1, x_1 = -1, x_2 = 1$
4. $(w_1 + w_2 - b) \geq 1,\ untuk\ y_4 = -1, x_1 = -1,\ x_2 = -1$

Selanjutnya jumlahkan persamaan

1. Jumlah persamaan (1) dan (2)

$$\begin{array}{r} (w_1 + w_2 + b) \geq 1 \\ (-w_1 + w_2 - b) \geq 1 \\ \hline 2w_2 = 2 \end{array} +$$

Maka $w_2 = 1$

2. Persamaan (1) dan (3)

$$\begin{array}{r} (w_1 + w_2 + b) \geq 1 \\ (w_1 - w_2 - b) \geq 1 \\ \hline 2w_1 = 2 \end{array} +$$

Maka $w_1 = 1$

3. Persamaan (2) dan (3)

$$\begin{array}{r} (-w_1 + w_2 - b) \geq 1 \\ (w_1 - w_2 - b) \geq 1 \\ \hline -2\mathrm{b} = 2 \end{array} +$$

Maka $\mathrm{b} = -1$

Sehingga persamaan hyperplanenya adalah

$$w_1 x_1 + w_2 x_2 + b = 0$$

$$w_1 + x_1 - 1 = 0$$

$$x_2 = 1 - x_1$$

SVM adalah salah satu metode klasifikasi ciri yang bertujuan menemukan hyperplane terbaik.

3.3.5 Analisis Hasil

Analisis hasil bertujuan untuk mengukur serta menarik kesimpulan, penelitian yang sedang dilakukan. Seberapa baik sistem menggambarkan proses dalam klasifikasi data. Pengukuran hasil akurasi pada penelitian ini dengan metode *Confusion matrix*. *Confusion matrix* adalah salah satu metode umum digunakan untuk melakukan perhitungan akurasi pada data mining[20]. Dari *Confusion matrix* akan dihitung yaitu akurasi, presisi dan recall.