Analisis Sentimen Terhadap Pembangunan Kereta Cepat Jakarta - Bandung Pada Media Sosial Twitter Menggunakan Metode SVM dan *GloVe Word Embedding*

1st Alam Rizki Fitriansyah
Fakultas Informatika
Universitas Telkom
Bandung, Indonesia
alamrfitriansyah@students.telkomuniversity.ac.id

2nd Yuliant Sibaroni
Fakultas Informatika
Universitas Telkom
Bandung, Indonesia
yuliant@telkomuniversity.ac.id

Abstrak-Proyek kereta cepat Jakarta – Bandung merupakan salah satu proyek besar yang saat ini sedang dibuat di Indonesia. Proyek kereta cepat Jakarta – Bandung menjadi ramai dibicarakan di media sosial Twitter, karena pada pembangunannnya terdapat banyak pihak yang merasa dirugikan, namun ada juga pihak yang merasa diuntungkan. Pada penelitian ini dilakukan analisis sentimen terhadap sentiment publik di media sosial Twitter tentang proyek kereta cepat Jakarta – Bandung. Penelitian ini menggunakan data yang berisi *tweet* dari keyword yang sudah ditentukan dan menggunakan GloVe *word embedding* dan metode klasifikasi *Support Vector machine*. Pada penelitian ini kombnasi terbaik pada parameter GloVe dengan nilai 200 untuk no_of_component, 0.001 untuk learning_rate dan fitur TOP 1 menghasilkan kenaikan pada nilai akurasi klasifikasi SVM dari 72.63% menjadi 77.72% dibandingkan dengan SVM tanpa menggunakan fitur ekspansi GloVe.

Kata kunci - sentimen, kereta cepat, twitter, GloVe, SVM

Abstract-The Jakarta – Bandung high-speed rail project is one of the major projects currently being built in Indonesia. The Jakarta – Bandung high-speed rail project has become a hot topic of discussion on Twitter social media, because during its construction there are many parties who feel disadvantaged, but there are also parties who feel benefited. In this study, sentiment analysis was carried out on public sentiment on Twitter social media about the Jakarta – Bandung high-speed rail project. This study uses data that contains tweets from predetermined keywords and uses GloVe word embedding and the Support Vector machine classification method. In this study, the best combination of GloVe parameters with a value of 200 for no_of_component, 0.001 for learning_rate and the TOP 1 feature resulted in an increase in the SVM classification accuracy value from 72.63% to 77.72% compared to SVM without using the GloVe expansion feature.

Keywords- sentiment, high-speed rail, twitter, GloVe, SVM

I. PENDAHULUAN

A. Latar Belakang

Pada masa kini media sosial memberikan peran dan pengaruh yang sangat besar dalam perkembangan teknologi komunikasi. Twitter merupakan salah satu media sosial microblogging yang memungkinkan penggunanya dapat membuat, melihat dan membalas postingan yang lebih dikenal dengan sebutan tweet. Pengguna Twitter sering kali menggunakan Twitter sebagai media untuk mengekspresikan diri dalam menanggapi suatu kejadian ataupun hal – hal yang terjadi di lingkungannya[1]. Hal ini dapat dijadikan sebagai acuan untuk mengetahui sentiment dan menentukan kecenderungan opini masyarakat terhadap suatu kejadian yang terjadi disekitar masyarakat. Salah satu kejadian yang ditanggapi oleh pengguna Twitter adalah proyek kereta cepat Jakarta – Bandung yang sudah dibangun sejak awal tahun 2016.

Proyek kereta cepat Jakarta – Bandung telah dimulai sejak tanggal 21 januari 2016 dengan dilakukannya groundbreaking oleh Jokowi di Perkebunan Mandalawangi Maswati, Cikalong Wetan, Bandung Barat, Jawa Barat[2]. Menurut Rini mantan menteri BUMN, keuntungan dibangunnya kereta cepat Jakarta – Bandung diantaranya akan meningkatkan perekonomian, mengangkat sektor pariwisata, dan membuka lapangan perkerjaan yang baru[3]. Namun, dalam pembangunannya terdapat beberapa masalah, seperti banjir yang terjadi di Bekasi dan menyebabkan kemacetan dan mengganggu kelancaran logistik. Permasalahan yang lain dalam proyek kereta cepat ini kurang memperhatikan kelancaran akses keluar – masuk jalan tol, pembiaran penumpukan material yang

mengganggu fungsi drainase, pembangunan pilar LRT tanpa izin, sampai persoalan keselamatan dan Kesehatan kerja (K3)[4]. Oleh karena itu, diperlukan analisis sentimen untuk mengetahui bagaimana sentimen yang ada pada media sosial mengenai pembangunan proyek kereta cepat Jakarta – Bandung.

Analisis sentiment adalah suatu cara untuk mendapatkan sebuah informasi sentiment pada suatu media sosial untuk memahami suatu opini atau preferensi dari pengguna media sosial. Informasi yang didapat bisa berupa sentimen yang bernilai positif atau negatif. Salah satu metode yang dapat digunakan untuk analisis sentiment adalah metode yang terdapat pada Machine Learning seperti, KNN, Naïve Bayes dan SVM. Tahapan – tahapan dasar yang dilakukan dalam menjalankan metode ini berupa crawling data, preprocessing, menambahkan label, menambahkan fitur, melakukan klasifikasi dan melakukan perhitungan sentimen. Seperti pada penelitian yang dilakukan oleh Pulung Hendro Prastyo menghasilkan tingkat rata – rata yang tinggi pada metode SVM untuk confusion matrix[6]. Namun, pada penelitian yang dilakukan oleh [5] Mohammad Rezwanul Huq menunjukan metode KNN memiliki hasil yang lebih tinggi dibandingkan dengan metode SVM pada penelitiannya. Oleh karena itu, jika hanya melihat dari tingkat akurasi, maka akan terjadi kekeliruan dalam penelitian, maka dalam melakukan penelitian perlu menambahkan precession atau recall sebagai evaluasi[5].

Dalam analisis sentimen dapat menggunakan word embedding sebagai fitur untuk merepresentasi kata berbentuk vektor bernilai rill. Pada penelitian yang dilakukan oleh [19] Liang-Chih Yu membahas tentang word embedding seperti, GloVe dan Word2vec pada analisis sentiment untuk menangkap sintaks dan semantic kata. Penelitian ini menggunakan word embedding pada klasifikasi Stanford Sentiment Treebank (SST) dan menghasilkan performansi yang lebih baik. Penelitian lainnya yang dilakukan oleh [20] Ru Ni dan Huan Cao memiliki hasil yang baik pada LTSM-GRU menggunakan GloVe sebesar 87.10% untuk akurasi dan 86.76% untuk F1-score.

Dengan adanya sentimen yang diungkapkan di media sosial *twitter* tentang proyek pembangunan kereta cepat Jakarta – Bandung akan dilakukan analisis sentimen menggunakan klasifikasi SVM dan GloVe *word embedding*. Oleh karena itu, Tugas Akhir ini akan mengangkat permasalahan mengenai pembangunan kereta Cepat Jakarta – Bandung dengan menggunakan metode klasifikasi Support Vector Machine (SVM). SVM memiliki kelebihan dalam menentukan hyperplane dengan memilih bidang yang optimal pada ruang input antar kelas dengan mencari margin yang merupakan titik data yang dekat dengan hyperplane[18]. Selain menggunakan SVM, penulis menggunakan ekspansi fitur Global Vector (GloVe) pada penelitian ini karena pada penelitian yang dilakukan oleh Ru Ni dan Liang-Chih Yu menghasilkan nilai performansi yang baik [19][20], serta menggunakan confucion matrix untuk melakukan pengukuran performansi.

B. Topik dan Batasan

Topik yang akan dibahas pada penelitian ini adalah analisis sentiment menggunakan metode klasifikasi *support vector machine* dan *Global Vector* untuk *word embedding* pada *tweet* masyarakat di media sosial *twittier* yang membahas seputar pembangunan kereta cepat Jakarta – Bandung. Batasan pada penelitian ini adalah menggunakan data set yang bersumber dari media sosial *twitter* dan hanya menggunakan kategori "proyek kereta cepat Jakarta – Bandung" dan data set yang diambil hanya menggunakan Bahasa Indonesia dengan format data .csv.

C. Tujuan

Tujuan dari penelitian ini adalah melakukan analisis sentiment pada media sosial twitter yang membahas tentang kereta cepat Jakarta – Bandung menggunakan metode klasifikasi *support vector machine* dan metode *word embedding global vector* untuk mengetahui performa model yang didapatkan dari metode SVM dan GloVe dengan menentukan kombinasi nilai parameter GloVe yang terbaik dan untuk mengetahui bagaimana sentiment publik mengenai proyek kereta cepat Jakarta – Bandung.

D. Organisasi Tulisan

Bagian dari laporan TA adalah sebagai berikut. Bab 2 yang membahas studi literatur yang terkait dengan penelitian ini. Kemudian pada bab 3 membahas sistem yang dibangun pada penelitian ini. Bab 4 membahas tentang hasil dan analisis dari penelitian dan bab 5 akan membahas kesimpulan dari penelitian yang sudah dilakukan.

II. KAJIAN TEORI

Twitter adalah sebuah media sosial yang menggabungkan format media jejaring sosial dan format blog atau yang disebut dengan 'microblogging'[10].Pada media sosial Twitter penggunanya menggunakan sebutan "Tweets" untuk sebuah konten yang diproduksi oleh akun profil lain. Twitter membatasi jumlah tweets sebanyak 140 karakter[12].Alasan mengapa Twitter banyak digunakan untuk melakukan analisis sentiment, karena penyebaran informasi pada Twitter terjadi sangat cepat sehingga memudahkan dalam mengetahui sentimen masyarakat terhadap suatu produk atau kejadian yang terjadi disekitarnya [13][14].

Analisis sentiment dapat dikatakan sebagai opinion mining yang bertujuan untuk menganalisa sebuah kalimat yang akan mendapatkan pendapat, penilaian, emosi seseorang, dan sentiment yang

berhubungan dengan suatu layanan, individu ataupun kegiatan lainnya[15].Analisis sentiment memiliki cara kerja dengan mengelompokan sebuah kalimat yang kemudian akan diberi nilai apakah kalimat tersebut bersifat positif, netral atau negatif dan analisis sentiment dapat menentukan sebuah perasaan emosional seperti senang atau sedih[16].

Dalam analisis sentiment terdapat berbagi macam fitur dan klasifikasi, salah satu ekspansi fitur yang digunakan adalah *Global Vector* (GloVe). GloVe akan menghasilkan output berupa *similarity words* untuk digunakan dalam melakukan ekspansi fitur[21]. Metode klasifikasi yang dapat digunakan salah satunya adalah metode klasifikasi *Support Vector Machine* (SVM) yang digunakan untuk mendeteksi teks[7]. Metode SVM merupakan sebuah algoritma yang menggunakan teori optimasi dna ruang hipotesis sebagai rangakai dalam bidang pattern recognition[1][17] dan menggunakan kernel untuk memetakan ruang input kedalam ruang fitur berdimensi tinggi.

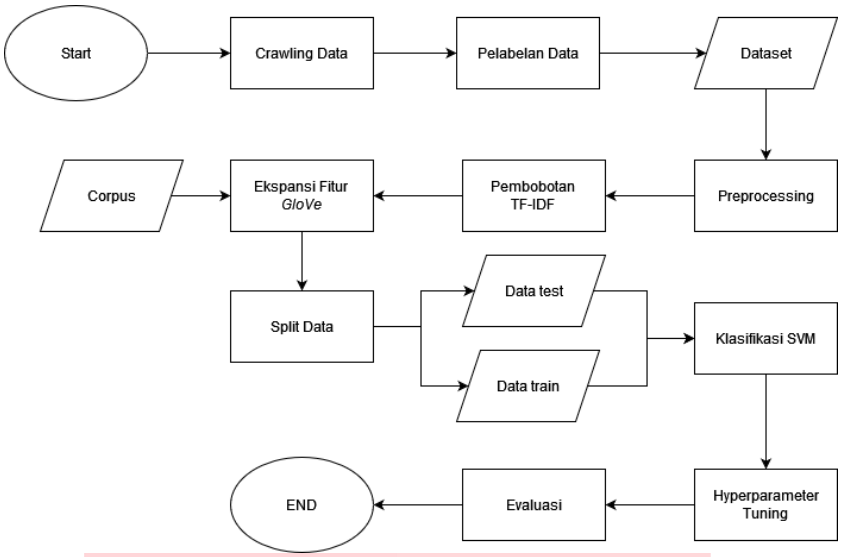
Pada penelitian yang dilakukan oleh Pulung Hendro Prastyo, metode SVM menghasilkan tingkat akurasi yang tinggi pada nilai rata – rata confusion matrix[6] dan pada penelitian yang dilakukan oleh [16] Angelina Puput Giovani, algortima SVM memiliki nilai akurasi dan performa yang paling tinggi dibandingkan dengan NB dan KNN. Pada tahun 2019[1] terdapat penelitian yang melakukan analisis sentiment terhadap bom bunuh diri Surabaya menggunakan metode Support Vector Machine (SVM). Data yang digunakan pada analisis sentiment ini adalah data yang berasal dari tweet yang di-posting oleh masyarakat karena tweet tersebut dapat merepresentasikan tanggapan masyarakat mengenai peristiwa tersebut. Analisis dilakukan dengan mengklasifikasi data menjadi 2 kondisi, yaitu sentiment positif sebanyak 121 tweets dan sentiment negatif sebanyak 1921 tweets. Dengan menggunakan metode SVM hasil klsifikasi pada data twitter tersebut memiliki tingkat akurasi sebesar 100% yang diuji menggunakan 1708 data testing dan 334 data training yang menunjukan bahwa secara keseluruhan klasifikasi menggunakan metode SVM menunjukan hasil yang baik pada data tweet tersebut. Penelitan lainnya yang menggunakan metode SVM dilakukan oleh [8] Kui Lu dan Jiesheng Wu melakukan analisis sentiment terhadap review film menggunakan metode SVM dan menghasilkan tingkat akurasi yang tinggi.

Penelitian lainnya[9] yang membahas tentang analisis sentimen dengan malakukan labeling menggunakan Vader dan ekstrasi menggunakan Term Frequency – Inverse Document Frequency (TF-IDF) yang pemodelan prediksi sentimen menggunakan Naïve Bayes dan Support Vector Machine (SVM). Data yang diambil merupakan data tweet sebanyak 15.494 tweet. Data yang digunakan untuk dijadikan data train sebanyak 10.845 data dan data yang dijadikan sebagai data test sebanyak 4649 data. Dengan menggunakan pemodelan Naïve Bayes dan Support Vector Machine (SVM) yrang menggunakan ekstrasi TF-IDF untuk memprediksi sentiment dari suatu tweets menghasilkan nilai akurasi sebesar 81% dengan f1-score 0.8 untuk Naïve Bayes sedangkan untuk Support Vector machine (SVM) menghasilkan nilai akurasi sebesar 87% dengan f1-score 0.87. Kemudian penelitian lainnya yang dilakukan oleh [20] Ru Ni dan Huan Cao menghasilkan tingkat akurasi sebesar 87.10% dan f1-score sebesar 86.76% pada model LSTM-GRU menggunakan GloVe. GloVe merupakan metode berbasis vektor yang pada umumnya menghasilkan nilai kesamaan yang tinggi dan dapat menjadi pilihan yang baik[21].

Berdasarkan penelitian terkait, algoritma pemodelan *Support Vector machine* (SVM) merupakan algoritma yang memiliki tingkat akurasi tinggi. Penggunaan GloVe sebagai word embedding menghasilkan performansi yang baik pada penelitian yang dilakukan oleh Ru Ni dan Liang-Chih Yu[19][20]. Pada penelitian ini akan menggunakan metode SVM untuk klasifikasi dengan ekspansi fitur menggunakan Global Vector (GloVe) untuk menghasilkan nilai akurasi dan performansi yang baik. Tujuan dari penelitian ini adalah melakukan analisis sentiment terhadap proyek kereta cepat Jakarta – Bandung menggunakan metode klasifikasi SVM pada media sosial Twitter.

III. METODE

Pada penelitian ini akan menggunakan metode Support Vector Machine dan Global Vector. Adapun skema pada penelitian ini yang dapat dilihat pada gambar 1.



GAMBAR 1
SKEMA UMUM

A. *Crawling* Data

Data set yang digunakan pada penelitian ini merupakan data set yang diambil dari media sosial Twitter dimulai tanggal tweet 1 Januari 2020 menggunakan beberapa keyword seperti #keretacepat, #dukungkeretacepat dan #keretacepatjakartabandung.

Pengumpulan data menggunakan metode crawling data menggunakan Twitter API yang telah tersedia pada library python seperti tweepy dan twint untuk mendapatkan data dari Twitter. Data yang didapatkan dari crawling data akan menjadi file dengan format Comma Separated Values (CSV) yang berisikan tweets berupa opini tentang proyek kereta cepat Jakarta – Bandung.

B. Pelabelan Data

Data yang telah terkumpul dilakukan pelabelan data secara manual dengan membagi menjadi 3 kelas/label yaitu positif, netral dan negatif. Pemberian label dilakukan oleh 3 orang kemudian memilih suara terbanyak dalam menentukan label dengan tujuan untuk mengurangi subjektivitas dalam melakukan pelabelan. Dalam pemberian label dilakukan dengan cara memperhatikan kata yang terdapat pada data tweets yang sudah diambil, apabila di dalam data tweets terdapat kata kasar atau kata yang tidak pantas makan akan diberi label negatif atau -1, apabila data tersebut tidak mengandung kata kasar dan memiliki kata yang bermakna positif maka diberikan label positif atau 1, dan data tersebut diberikan label netral jika kalimat tersebut tidak mengandung makna positif atau negatif.

Kata kasar memiliki artian umpatan, cacian, cercaan, ejekan, dan lainnya. Sedangkan kalimat yang kurang pantas adalah kalimat yang mengandung provokasi yang dapat berdampak buru bagi yang membacanya atau kalimat negatif yang dapat memperburuk keadaan.

TABEL 1
CONTOH PELABELAN DATA

Label	Tweet
Positif	Pemerintah menjamin keselamatan masyarakat dengan menerapkan SOP rancang bangun proyek Kereta Cepat Jakarta-Bandung yang profesional
Netral	Proses pengerjaan kereta cepat Jakarta Bandung diikuti dengan transfer teknologi. #VideoBerita #AdadiKompas
Negatif	Pantess salah pengerjaan..emang da ada harusnya dr awal perencanaan proyek…yg ada nambah utang..ruwett ruwettt..yg ga setuju lsg diganti, kursi diisi politik balas budi dgn org2 yg ga kompeten..ambyarrrr #keretacepat

C. *Preprocessing* Data

Preprocessing memiliki tujuan untuk membersihkan dan mengubah data yang tidak diperlukan dalam melakukan klasifikasi agar memiliki hasil yang lebih optimal. Tahapan Preprocessing dalam penelitian ini terdapat beberapa tahap, yaitu :

1. *Data Cleaning*

Data cleaning merupakan tahap untuk membersihkan data dan memperbaiki inkonsistensi dalam data dengan cara menghapus tanda baca, *hashtag*, *URL*, symbol, angka dan atribut yang kosong. Contoh

untuk yang akan terhapus ada pada tabel 2 dan contoh untuk proses *data cleaning* seperti pada tabel 3.

TABEL 2
CONTOH TANDA BACA, SIMBOL DAN ANGKA

(.) , (,) , (?) , (!) , (;) , (:) , (-) , (--) , (=), (‘..’) , (“..”) , (/) , ((..)) , ([..]) , (`) , (~), (@) , (#) , ($) , (^) , (&) , (*) , (_) , (+), ({..}) . (|) , (>) , (<) , (1,2,3,4,5,6,7,8,9)

TABEL 3
DATA CLEANING

Sebelum	Sesudah
Mari kita Dukung Proyek Kereta Cepat yg akan berikan manfaat utk pertumbuhan pemerataan ekonomi Indonesia. **#DukungKeretaCepat**	Mari kita Dukung Proyek Kereta Cepat yg akan berikan manfaat utk pertumbuhan pemerataan ekonomi Indonesia.

2. *Case Folding*

Case folding adalah proses perubahan huruf dari awalan kata yang sebelumnya huruf besar menjadi huruf kecil. Contoh untuk *case folding* seperti pada tabel 4.

TABEL 4
CASE FOLDING

Sebelum	Sesudah
Mari kita **Dukung Proyek Kereta Cepat** yang akan berikan manfaat untuk pertumbuhan pemerataan ekonomi **Indonesia**.	**mari** kita **dukung proyek kereta cepat** yang akan berikan manfaat untuk pertumbuhan pemerataan ekonomi **indonesia**.

3. *Normalization*

Normalization merupakan tahap untuk melakukan identifikasi terhadap penulisan kata yang berlebihan kemudian akan dilakukan penggantian kata dengan yang sesuai dengan KBBI menggunakan kamus kata yang dibuat secara manual dan juga menggunakan library Sastrawi. Contoh untuk *normalization* seperti pada tabel 5.

TABEL 5
NORMALIZATION

Sebelum	Sesudah
mari kita dukung proyek kereta cepat yg akan berikan manfaat utk pertumbuhan pemerataan ekonomi Indonesia.	mari kita dukung proyek kereta cepat yang akan berikan manfaat untuk pertumbuhan pemerataan ekonomi Indonesia.

4. *Stop Word Removal*

Stop Word Removal merupakan porses *filtering* terhadap kata yang dianggap tidak penting atau kata umum yang mempunyai fungsi namun tidak memiliki arti. Proses penghapusan *stopword* dalam Bahasa Indonesia menggunakan kamus kata yang dibuat secara manual dan juga menggunakan *library* Sastrawi. Contoh untuk *stop word removal* seperti pada tabel 6.

TABEL 6
STOP WORD REMOVAL

Sebelum	Sesudah
mari kita dukung proyek kereta cepat yang akan berikan manfaat untuk pertumbuhan pemerataan ekonomi Indonesia.	mari kita dukung proyek kereta cepat berikan manfaat pertumbuhan pemerataan ekonomi indonesia.

5. *Stemming*

Stemming merupakan proses untuk menghilangkan kata dan mengubah kata imbuhan menjadi kata dasar dengan cara menghapus imbuhan yang terdapat pada kata tersebut. Proses *stemming* menggunakan *library* pada *python* yang khusus untuk Bahasa Indonesia yaitu *library* Sastrawi. Contoh untuk *stemming* seperti pada tabel 7.

TABEL 7
STEMMING

Sebelum	Sesudah
mari kita dukung proyek kereta cepat berikan manfaat pertumbuhan pemerataan ekonomi	mari kita dukung proyek cepat berikan tumbuh rata ekonomi indonesia

Indonesia.	

6. Tokenization

Tokenization adalah proses memisahkan kalimat menjadi kata yang sebelumnya dipisahkan oleh spasi yang dilakukan untuk mempermudah klasifikasi. Contoh untuk *Tokenization* seperti pada tabel 8.

TABEL 8
TOKENIZATION

Sebelum	Sesudah
mari kita dukung pyotek cepat berikan tumbuh rata ekonomi indonesia	"mari", "kita", "dukung", "proyek", "cepat", "berikan", "tumbuh", "rata", "ekonomi", "indonesia"

C. Pembobotan Term Frequency – Inversce Document Frequency (TF-IDF)

Pada penelitian ini, ekstraksi fitur yang digunakan yaitu Term Frequency – inverse document frequency (TF-IDF). TF-IDF adalah sebuah metode yang mengitegrasikan antara Term Frequency (TF) yang menghitung jumlah kata yang muncul dari kalimat dengan Inverse Focument Frequency (IDF) yang menunjukan seberapa sering kata tersebut muncul dalam sebuah dokumen[20]. TF-IDF berfungsi untuk mencari representasi nilai dari tiap tiap dokumen dari kumpulan data training yang dibentuk menjadi vektor kemudian akan dicari kesamaan antar dokumen.

.

TF-IDF memliki persamaan sebagai berikut :

$$tf_i = \frac{freq_i(d_j)}{\sum_{i=1}^{k} freq_i(d_j)} \quad (1)$$

$$idf_i = \log \frac{|D|}{|\{d:t_i \in d\}|} \quad (2)$$

$$(tf - idf)_{ij} = tf_i(d_j) * idf_i \quad (3)$$

Pada persamaan tf_i merupakan perhitungan dari kata yang dicari dalam sebuah kalimat dengan cara menghitung kemunculan kata yang dicari pada suatu kalimat. idf_i merupakan *inversed document frequency*, dimana nilai idf didapatkan dari nilai *log* dari total dokumen dibagi dengan banyak kata yang dicari dalam dokumen, lalu tf_i dikalikan dengan . idf_i.

D. Global Vector (GloVe)

Ekspansi fitur yang digunakan pada penelitian ini adalah Global Vector (GloVe). GloVe memiliki cara kerja dengan membuat matriks yang akan menghitung seberapa sering sebuah kata muncul. GloVe akan menghasilkan output list yang berisikan similarity words yang selanjutnya akan dilanjutkan dengan proses ekspansi terhadap vektor[21].

Glove memiliki perhitungan - perhitungan sebagai berikut :

$$w_i^T + \vec{w}_k + b_i + \vec{b}_k = \log(X_{i_k}) \quad (4)$$

Di mana :
w = vektor kata
$\vec{w}$ = Vektor kontek kata
b_i = bias skalar kata utama
b_k = bias skalar kontek kata
X = matriks kemunculan
Perhitungan untuk menghitung nilai f (Xik) dilakukan dengan persamaan dibawah ini :

$$f(X_{ik}) = \begin{cases} \left(\frac{X_{i_k}}{xmax}\right)^a ; i f X_{i_k} < xmax \\ 1 ; lainnya \end{cases} \quad (5)$$

Dari persamaan diatas didapatkan fungsi pembobotan ke dalam fungsi cost yang memberikan persamaan sebagai berikut :

$$J = \sum_{i,k=1}^{v} f(X_{i_k}) \left(w_i^T \vec{w} j + b_i + b_k - \log X_{i_k}\right)^2 \quad (6)$$

Contoh dari penerapan *word embedding* GloVe dijabarkan pada tabel 9.

TABEL 9
CONTOH DOKUMEN

D1	Kereta cepat jakarta bandung
D2	Kereta cepat hanya sampai jakarta

TABEL 10
GLOVE

	Kereta	cepat	jakarta	bandung	hanya	Sampai	padalarang
Kereta	0	2	0	0	0	0	0
Cepat	2	0	1	0	1	0	0
Jakarta	0	1	0	1	0	0	0

Bandung	0	0	1	0	0	0	0
Hanya	0	1	0	0	0	1	0
Sampai	0	0	0	0	1	0	1
Padalarang	0	0	0	0	0	1	0

Pada penelitian ini akan menggunakan library yang sudah tersedia pada python dalam library glove-pyhton. Sebelum melakukan ekspansi menggunakan metode GloVe akan dilakukan preprocessing, seperti data cleaning, data normalization, stop word removal, dan lainnya. Untuk membuat model GloVe diperlukan co-occurrence matrix yang dihasilkan oleh corpus object untuk membuat embedding[22].

Terdapat dua paramater dari GloVe, yaitu no_of_components yang menunjukan dimensi dari output vector yang dihasilkan oleh GloVe dan learning_rate untuk menentukan kecepatan dari algoritma[23]. Algoritma GloVe akan menghasilkan *output* berupa list dari *similarity words* seperti pada tabel 11 dengan kata "kereta".

TABEL 11
CONTOH SIMILARITY WORD

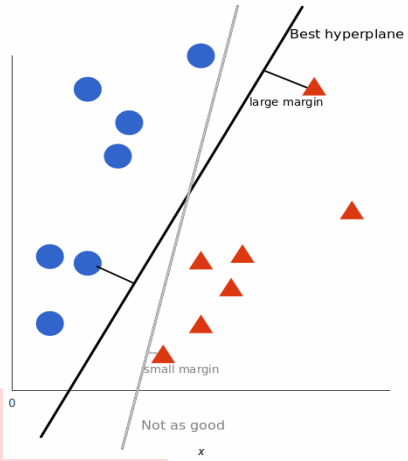
Kata	**Rank 1**	**Rank 2**	**Rank 3**	**Rank 4**	**Rank 5**
Kereta	Contoh	Semi	Arti	Jalur	Kayak
	Rank 6	**Rank 7**	**Rank 8**	**Rank 9**	**Rank 10**
	Dengan	Tarif	Regulasi	Layang	Nama

E. Klasifikasi Support Vector Machine (SVM)

Sebelum melakukan klasifikasi, dilakukan pembagian data atau *split data* dengan rasio perbandingan 90:10, 90% data train dan 10% data test. Kemudian pada penelitian ini klasifikasi dilakukan dengan menggunakan metode Support Vector Machine (SVM). SVM adalah sebuah algoritma pembelajaran yang menggunakan teori optimasi dan ruang hipotesis dalam sebuah fitur yang berdimensi tinggi yang dapat digunakan untuk klasifikasi dan analisis regresi. Algoritma SVM diperkenalkan pada tahun 1992 oleh Vapnik sebagai rangkaian dalam bidang pattern recognition[1][17]. SVM memiliki tujuan untuk menemukan hyperplane pemisah maksimum yang optimal pada ruang input antar kelas dengan mencari margin yang merupakan titik data yang dekat dengan hyperplane[18]. Persamaan hyperplane klasifikasi SVM dituliskan dengan :

$$\vec{w}.\ \vec{x} + b = 0 \qquad (7)$$

Di mana :
$\boldsymbol{w}$ = parameter bobot,
$\boldsymbol{x}$ = vector input,
$\boldsymbol{b}$ = bias.



GAMBAR 2
SVM

pada hyperplane terdapat dua kelas yaitu kelas "naik" dan "turun" dengan garis vektor yang dapat diberikan oleh [1]:

$$\vec{w}.\ \vec{x} + b \geq 1, untuk\ Y1 = +1 \qquad (8)$$
$$\vec{w}.\ \vec{x} + b \leq 1, untuk\ Y1 = -1 \qquad (9)$$

Sampel yang memiliki nilai negatif akan masuk kedalam kelas -1, sedangkan sampel yang miliki nilai positif akan masuk ke dalam kelas +1.

F. Hyperparameter Tuning

Hyperparameter Tuning adalah tahapan untuk menghasilkan model yang optimal dan meningkatkan kinerja model yang disesuaikan dengan pemodelan tertentu. Pada penelitian ini, hyperparameter dimulai dengan melakukan inisialisasi hyperparameter yang dioptimalkan kemudian menggunakan grid search untuk menentukan hyperparameter terbaik.

Grid search menguji coba kombinasi – kombinasi yang memiliki tujuan untuk menentukan kombinasi parameter untuk SVM yang memilki hasil performa model terbaik dengan menghitung nilai rata – rata cross validation untuk setiap kombinasi. Hyperparameter yang dapat dioptimalkan dalam SVM adalah parameter C dan parameter gamma. Parameter C bekerja dengan cara mengoptimalkan SVM dan menghindari terjadinya misklasifikasi didalam data train, sedangkan parameter gamma bertugas untuk menentukan pengaruh pada data train.

Fungsi kernel pada SVM yang digunakan pada *hyperparameter tuning* adalah sebagai berikut[25] :

1. Linear :
$$\langle x, x' \rangle \qquad (10)$$

2. Polinomial :
$$(\gamma \langle x, x' \rangle + r)^d \qquad (11)$$

dimana d diartikan sebagai parameter *degree* dan r adalah coef0

3. RBF :

$$\exp(-\gamma\|x - x'\|^2) \quad (12)$$

dimana γ diartikan sebagai parameter gamma.

4. Sigmoid :

$$\tanh(\gamma\langle x, x'\rangle + r) \quad (13)$$

dimana r diartikan sebagai coef0.

G. Evaluasi *Confusion Matrix*

Confusion matrix adalah matriks untuk menampilkan hasil prediksi dengan hasil aktual.

TABEL 12
CONFUSION MATRIX

Kelas	**Prediksi**	
	Positif	**Negatif**
Positive	*True Positive* (TP)	*False Negative* (FN)
Negative	*False Positive* (FP)	*True Negative* (TN)

Parameter yang terdapat pada tabel 12 dapat digunakan untuk menghitung nilai performansi klasifikasi seperti accuracy, precission, recall, specifity dan f1-score.

1. Accuracy

Accuracy adalah nilai yang menunjukan tingkat ketepatan dari prediksi model yang dapat mengklasifikasikan dengan benar dibandingkan dengan data yang digunakan. Accuracy memiliki persamaan sebagai berikut :

$$\text{Accuracy} = \frac{(TP+TN)}{(TP+TN+FP+FN} \quad (14)$$

2. Precission

Precission adalah perbandingan antara data positif dengen keseluruhan hasil prediksi. Precission memiliki persamaan sebagai berikut :

$$\text{Precission} = \frac{TP}{(TP+FP)} \quad (15)$$

3. Recal

Recall adalah tingkat perbandingan antara data yang memiliki nilai positif yang diprediksi benar dengan keseluruhan data yang diprediksi positif benar ataupun tidak. Recall memiliki persamaan sebagai berikut :

$$\text{Recall} = \frac{TP}{(TP+FN)} \quad (16)$$

4. F1-Score

f1-score merupakan perbandingan nilai antara precission dan recall. F1-score memiliki perbandingan sebagai berikut :

$$\text{f1-score} = \frac{Precission\ X\ Recall}{(Precission+Recall)} \quad (17)$$

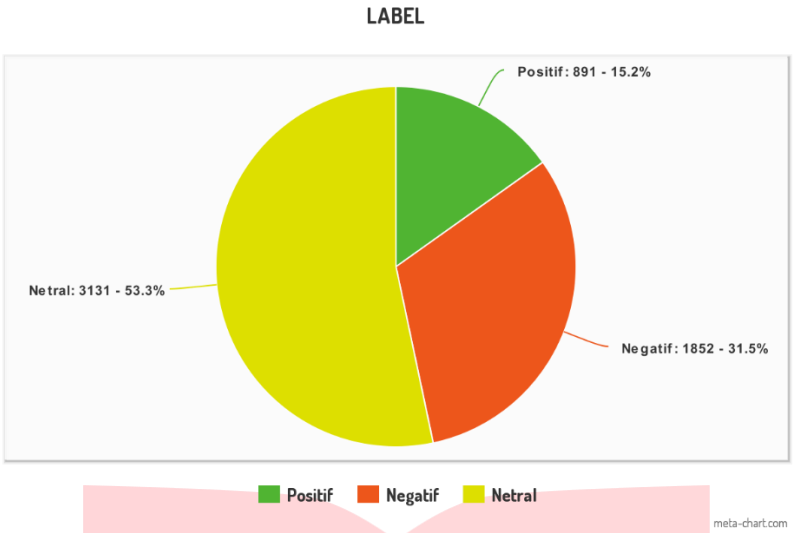
IV. HASIL DAN PEMBAHASAN

A. Data

Data *tweet* yang digunakan sebanyak 5.874 *tweet* berbahasa Indonesia dengan topik pembangunan kereta cepat Jakarta – Bandung yang didapatkan dari beberapa keyword yaitu #keretacepat, #dukungkeretacepat dan #keretacepatjakartabandung. Data yang digunakan dibagi menjadi 3 label yaitu label positif, negatif, dan netral. dengan rincian seperti pada tabel 13 dan gambar 3.

TABEL 13
PERSEBARAN DATA

Sentimen	**Jumlah data**
Positif	891
Netral	3131
Negatif	1852



GAMBAR 3
PERSEBARAN DATA

Label 1 diberikan untuk sentiment positif, label 0 diberikan untuk sentiment netral, dan label -1 diberikan untuk sentimen negatif.

B. Pembuatan Corpus

Corpus atau kamus kata yang digunakan pada penelitian ini berasal dari data *twitter* dengan menggunakan metode *Global Vector* dengan ukuran window 3 dan didapatkan dari kumpulan kata yang berasal dari dataset sebanyak 5.875 data *tweet* sehingga mendapatkan 8.733 kosakata pada *corpus twitter*.

C. Pembobotan TF-IDF

Pada proses pembobotan TF-IDF dilakukan pemberian bobot kata pada setiap kalimat yang ada pada data yang dimaksud dengan tujuan untuk mengetahui nilai kemunculan kata pada dokumen tersebut menggunakan persamaan (1), (2) dan (3).

D. Skenario dan Hasil Uji

Pada penelitian ini dilakukan 2 skenario, yaitu :

1. Pengujian performansi menggunakan GloVe pada model klasifikasi SVM
2. Pengujian performansi menggunakan *hyperparameter tuning*.

Sebelum dilakukan skenario pertama, yang pertama dilakukan adalah melakukan klasifikasi *Support Vector Machine* tanpa menggunakan fitur ekspansi *Global Vector* untuk digunakan sebagai baseline pada penelitian ini dengan tujuan untuk mengetahui pengaruh dari adanya penambahan proses fitur ekspansi GloVe dan klasifikasi SVM tanpa menggunakan GloVe menghasilkan tingkat akurasi sebesar 72.63%

1. Skenario Satu

Skenario yang pertama adalah melakukan pengujian performansi menggunakan GloVe pada model klasifikasi SVM. Pengujian ini dilakukan dengan cara mencari kombinasi nilai parameter GloVe yaitu no_of_components dan learning_rate yang memberikan tingkat akurasi tertinggi terhadap klasifikasi SVM kemudian akan ditambahkan dengan fitur *similarity word* (TOP), karena pada tingkat akurasi dapat menunjukan kedekatan antara hasil pengukuran klasifikasi dengan nilai yang sesungguhnya. No_of_components menunjukan dimensi dari output vector pada tiap kata yang terdapat pada corpus dan learning_rate untuk menentukan kecepatan dari algoritma. Sehingga didapatkan kombinasi dan tingkat akurasi seperti pada tabel 14.

TABEL 14
HASIL PERFORMANSI SVM + GLOVE (TOP 1)

Akurasi (%)				
No_of_component	**Learning_rate**			
	0.001	**0.005**	**0.01**	**0.05**
50	75,96 %	73,33 %	74,73 %	73,33 %
100	76,84 %	72,63 %	77,37 %	75,78 %
200	**77,72 %**	75,29 %	77,19 %	75,43 %
300	76,84 %	75,43 %	75,26 %	73,85 %

Berdasarkan hasil yang didapatkan pada tabel 14, didapatkan kombinasi terbaik dengan nilai parameter 200 untuk no_of_components, nilai parameter 0.001 untuk learning rate dan menggunakan fitur TOP 1 dengan nilai akurasi pada klasifikasi SVM sebesar 77.72%, sehingga kombinasi nilai parameter tersebut merupakan kombinasi yang terbaik untuk penelitian ini dan kombinasi tersebut dilanjutkan dengan melakukan uji terhadap fitur TOP 3, TOP 5, TOP 7 dan TOP 9. Hasil uji terhadap fitur TOP 3, TOP 5, TOP 7 dan TOP 9 dapat dilihat pada tabel 15.

TABEL 15
HASIL PERFORMANSI FITUR GLOVE

Fitur	Akurasi
Top 3	77,54%
Top 5	77,02%
Top 7	76,14%
Top 9	75,61%

2. Skenario Dua

Skenario kedua adalah melakukan pengujian performansi menggunakan *hyperparameter tuning* dengan menentukan parameter yang terbaik pada hyperparameter tuning dan didapatkan anatara model liner default dan *hyperparameter tuning* tidak memeberikan hasil yang berbeda, yaitu nilai akurasi sebesar 77.72% pada kernel RBF, nilai 10 untuk C dan 0.1 untuk gamma.

E. Analisis Hasil Pengujian

Berdasarkan hasil dari pengujian yang dilakukan, pengujian dilakukan pada nilai parameter no_of_component dan learning_rate dapat memberikan hasil yang berbeda – beda tergantung daripada ukuran dimensi corpus dan kecepatan algoritma yang ditentukan pada kedua parameter tersebut. Dari pengujian yang telah dilakukan, proses fitur ekspansi GloVe menunjukan hasil yang baik pada tingkat akurasi klasifikasi SVM, kenaikan dari 72.63% menjadi 77.72% dan pada dataset yang digunakan pada penelitian ini, nilai kombinasi yang terbaik untuk no_of_component adalah 200, untuk learning_rate adalah 0.001 dan menggunakan fitur TOP 1. Hal tersebut dikarenakan fitur TOP 1 lebih spesifik dalam mencari tingkat similaritas terhadap setiap kata.

V. KESIMPULAN

Berdasarkan penelitian analisis sentimen dengan menggunakan GloVe dan klasifikasi *Support Vector Machine* yang dilakukan ini dapat disimpulkan bahwa GloVe dapat memberikan hasil yang baik kepada tingkat akurasi dari klasifikasi *Support Vector Machine* dengan adanya peningkatan nilai akurasi dari 72.63% menjadi 77.72%. Pada penelitian ini kombinasi dari parameter GloVe dapat mempengaruhi hasil dari tingkat akurasi *Support Vectore Machine* dan pada penelitian ini nilai kombinasi terbaik untuk parameter GloVe adalah 200 untuk no_of_component, 0.001 untuk learning_rate dam fitur TOP 1. Parameter no_of_component berpengaruh terhadap ukuran vektor dimensi pada setiap kata yang terdapat pada *Corpus twitter* dan parameter learning_rate untuk menentukan kecepatan dari algoritma. Kemudian untuk *Hyperparameter tuning* pada penelitian ini tidak memberikan hasil yang berbeda dengan model default.

Saran untuk penelitian selanjutnya adalah mencoba menggunakan teknik imbalance untuk mengetahui apakah teknik imbalance dapat mempengaruhi nilai akurasi pada algoritma klasifikasi.

REFERENSI

[1] Ihsan, M., Roza, E., & Widodo, E. (2019). Analisis Sentimen Twitter terhadap Bom Bunuh Diri di Surabaya 13 Mei 2018 menggunakan Pendekatan Support Vector Machine. PRISMA, Prosiding Seminar Nasional Matematika, 2, 416–426.

[2] PPID. (2016). Presiden Jokowi Resmikan Groundbreaking Proyek Kereta Cepat Dan Sentra Ekonomi Koridor Jakarta -Band. Diakses pada 25 Oktober 2021, dari http://ppid.menlhk.go.id/berita/berita-foto/340/presiden-jokowi-resmikan-groundbreaking-proyek-kereta-cepat-dan-sentra-ekonomi-koridor-jakarta-band

[3] Handyono. (2016). Manfaat dari proyek kereta cepat Jakarta – Bandung. Diakses pada 31 Oktober 2021, dari 3

[4] Gusman, Hanif. (2020). Fakta dan Masalah Kereta Cepat Jakarta – Bandung. https://tirto.id/fakta-dan-masalah-kereta-cepat-jakarta-bandung-eG7s

[5] Rezwanul, M., Ali, A. and Rahman, A. (2017) 'Sentiment Analysis on Twitter Data using KNN and SVM', International Journal of Advanced Computer Science and Applications, 8(6), pp. 19–25. doi: 10.14569/ijacsa.2017.080603.

[6] Prastyo, P. H. et al. (2020) 'Tweets Responding to the Indonesian Government's Handling of COVID-19: Sentiment Analysis Using SVM with Normalized Poly Kernel', Journal of Information Systems Engineering and Business Intelligence, 6(2), p. 112. doi: 10.20473/jisebi.6.2.112-122..

[7] Ahmad, M., Aftab, S. and Ali, I. (2017) 'Sentiment Analysis of Tweets using SVM', International Journal of Computer Applications, 177(5), pp. 25–29. doi: 10.5120/ijca2017915758.

[8] Lu, K. and Wu, J. (2019) 'Sentiment analysis of film review texts based on sentiment dictionary and SVM', PervasiveHealth: Pervasive Computing Technologies for Healthcare, Part F148152, pp. 73–77. doi: 10.1145/3319921.3319966.

[9] Alizah, M. D., Nugroho, A., Radiyah, U., & Gata, W. (2020). Sentimen Analisis Terkait Lockdown pada Sosial Media Twitter. Indonesian Journal on Software Engineering (IJSE), 6(2), 223–229. https://doi.org/10.31294/ijse.v6i2.8991

[10] Rumata, Vience Mutriara. (2017). Analisis Isi Kualitatif Twitter "#TaxAmensty" dan "#AmenestiPajak". PIKOM, Penelitian Komunikasi dan Pembangunan.

[11] Nurhadi, Z. F. (2017). Model Komunikasi Sosial Remaja Melalui Media Twitter. Jurnal

ASPIKOM, 3(3), 539. https://doi.org/10.24329/aspikom.v3i3.154

[12] Akbar, M. T., Martutik, M., & Safii, M. (2018). Konten Akun Media Sosial Twitter Perpustakaan Universitas Perguruan Tinggi Di Indonesia. BIBLIOTIKA : Jurnal Kajian Perpustakaan Dan Informasi, 2(1), 41–49. https://doi.org/10.17977/um008v2i12018p041

[13] Rustiana, D., & Rahayu, N. (2017). Analisis Sentimen Pasar Otomotif Mobil: Tweet Twitter Menggunakan Naïve Bayes. Simetris: Jurnal Teknik Mesin, Elektro Dan Ilmu Komputer, 8(1), 113–120. https://doi.org/10.24176/simet.v8i1.841

[14] Liu, B. (2010). Sentiment analysis and subjectivity. In Handbook of Natural Language Processing, Second Edition.

[15] deHaaff, Michelle. (2010). Sentiment Analysis, Hard But Wort It!. Diakses pada 23 November 2021, dari https://customerthink.com/sentiment_analysis_hard_but_worth_it/

[16] Giovani, A. P., Ardiansyah, A., Haryanti, T., Kurniawati, L., & Gata, W. (2020). Analisis Sentimen Aplikasi Ruang Guru Di Twitter Menggunakan Algoritma Klasifikasi. Jurnal Teknoinfo, 14(2), 115. https://doi.org/10.33365/jti.v14i2.679

[17] T, Y. S., Faraby, S. Al, & Mahendra Dwifebri. (2019). Analisis Sentimen Terhadap Ulasan Film Menggunakan Word2Vec dan SVM. 8(4), 4136–4144.

[18] A. S. Akbar, E. Sediyono, and O. D. Nurhayati, "Analisis Sentimen Berbasis Ontologi di Level Kalimat untuk Mengukur Persepsi Produk," J. Sist. Inf. Bisnis, vol. 5, no. 2, pp. 84–97, 2015, doi: 10.21456/vol5iss2pp84- 97.

[19] L. C. Yu, J. Wang, K. R. Lai, and X. Zhang, "Refining word embeddings for sentiment analysis," EMNLP 2017 - Conf. Empir. Methods Nat. Lang. Process. Proc., pp. 534–539, 2017, doi: 10.18653/v1/d17-1056.

[20] Widyasanti, N., Darma Putra, I. and Dwi Rusjayanthi, N., 2018. Seleksi Fitur Bobot Kata dengan Metode TFIDF untuk Ringkasan Bahasa Indonesia. Jurnal Ilmiah Merpati (Menara Penelitian Akademika Teknologi Informasi), Vol 6(2252-3006), p.119.

[21] R. Ni and H. Cao, "Sentiment Analysis based on GloVe and LSTM-GRU," 2020 39th Chinese Control Conference (CCC), 2020, pp. 7492-7497, doi: 10.23919/CCC50068.2020.9188578.

[22] R. D. Indrapurasih, M. A. Bijaksana, I. L. Sardi, and L. Belakang, "Implementasi dan Analisis Kesamaan Semantik Antar Kata Bahasa Indonesia Menggunakan Metode GloVe Pendahuluan Studi Terkait Semantic Similarity," eProceedings Eng., vol. 5, no. 3, pp. 7699–7706, 2018.

[23] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global vectors for word representation," EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference. pp. 1532–1543, 2014, doi: 10.3115/v1/d14-1162.

[24] J. S. Chawla. (2018). What is Glove?. Diakses pada 12 Januari 2022, dari https://medium.com/analytics-vidhya/word-vectorization-using-glove-76919685ee0b

[25] Scikit-learn developers. Support Vector Machines. Diakses pada 23 Juli 2022, dari https://scikit-learn.org/stable/modules/svm.html