JMPM: Jurnal Matematika dan Pendidikan Matematika
9 (1) Maret 2024 1-14
Tersedia secara online di https://www.journal.unipdu.ac.id/index.php/jmpm
Terakreditasi S3 – SK No. B/1796/E5.2/KI.02.00/2020

Implementasi Metode *Naïve Bayes* dan *Support Vector Machine (SVM)* untuk Menganalisis Sentimen Pengguna Twitter terhadap Transjakarta

(*Implementation of Naïve Bayes and SVM Methods to Analyze Twitter User Sentiment on Transjakarta*)

Khairunnisa Fadhilla Ramdhania[1*], Dian Fitrianto Hidayat[2], Ratna Salkiawati[3]

[1,2,3] Program Studi Informatika, Universitas Bhayangkara Jakarta Raya– Jl. Raya Perjuangan, Marga Mulya, Bekasi Utara, 17123

[*] email penulis korespondensi: khairunnisa.fadhilla@dsn.ubharajaya.ac.id

Abstrak

Twitter adalah media sosial populer di Indonesia yang digunakan untuk mengungkapkan perasaan dan menyampaikan opini. Transjakarta adalah transportasi umum dengan jumlah penumpang harian terbanyak. Pengguna Transjakarta kerap berbagi pengalaman baik atau buruk, serta opini di Twitter. Beberapa masalah pada Transjakarta yang dibahas di Twitter yaitu terkait penutupan halte sementaradan kasus asusila yang dialami oleh pengguna Transjakarta. Artikel ini menganalisis opini pengguna Twitter tentang Transjakarta menggunakan metode *Support Vector Machine (SVM*) dan *Naïve Bayes*. *Lexicon based* digunakan untuk pemberian label pada 6736 tweet. Hasil analisis menunjukkan 2228 tweet positif dan 2821 tweet negative. Metode *Support Vector Machine* mencapai akurasi 84.95%, presisi 83%, *recall* 83% dan *f1-score* 83%, sedangkan *Naïve Bayes* mencapai akurasi 76.43%, presisi 78%, *recall* 68% dan *f1-score* 73%.

Kata kunci: *twitter, transjakarta, lexicon based, naïve bayes, support vector machine*

Abstract

Twitter is a popular social media platform in Indonesia that allows people to express their feelings and share their opinions. Transjakarta is the public transportation system with the highest number of daily passengers. Transjakarta users often share their positive and negative experiences on Twitter. Some issues discussed on Twitter regarding Transjakarta include temporary bus stop closures and indecent incidents. This article analyzes Twitter users' opinions about Transjakarta using the Support Vector Machine (SVM) and Naïve Bayes methods. A lexicon-based approach was used to label 6,736 tweets. The analysis results show 2,228 positive tweets and 2,821 negative tweets. The SVM method achieved an accuracy of 84.95%, precision of 83%, recall of 83%, and an F1-score of 83%, while Naïve Bayes achieved an an accuracy of 76.43%, precision of 78%, recall of 68%, and an F1-score of 73%.

Keywords: *twitter, transjakarta lexicon based, naïve bayes, support vector machine.*

Cara mengutip dengan APA 7 style: Ramdhania, K. F., Hidayat, D. F., & Salkiawati, R. (2024). Implementasi metode naïve bayes dan svm untuk menganalisis sentimen pengguna twitter terhadap transjakarta. *JMPM: Jurnal Matematika dan Pendidikan Matematika*, *9*(1), 1-14. https://dx.doi.org/10.26594/jmpm.v9i1.4494.

PENDAHULUAN

Perkembangan teknologi internet yang semakin pesat telah berdampak pada cara antarindividu dan kelompok dalam melakukan komunikasi. Hadirnya internet menjadi peluang munculnya berbagai platform digital. Berkaitan dengan hal tersebut, media sosial termasuk *platform* yang mengambil peran untuk dapat berbagai informasi, berinteraksi, dan bekerja sama antar individu maupun kelompok. Twitter merupakan dari salah satu media sosial yang populer di Indonesia. Twitter dapat berguna untuk mengungkapkan perasaan bagi para penggunanya dan memberikan kritik atau saran terhadap apapun.

Berdasarkan laporan *We Are Social* dan *Hootsuite* (Annur, 2023), terdapat 556 juta pengguna Twitter di seluruh dunia pada Januari 2023. Jumlah tersebut meningkat 27,4% dibandingkan pada periode yang sama tahun sebelumnya. Meski mempunyai banyak pengguna secara global, namun menurut laporan *We Are Social*, Twitter menempati peringkat ke-14 aplikasi media sosial dengan jumlah pengguna terbanyak. Indonesia menempati peringkat lima dengan jumlah pengguna yang mencapai 24 juta.

Pengguna Twitter membicarakan suatu isu dengan berbagai cara, yang dapat dilihat dari aktivitas di *platform* tersebut. Cara yang paling umum untuk menyuarakan pendapat pada Twitter adalah dengan menuliskan *tweet* dan melakukan *retweet*. Isi *tweet* dapat dijadikan sebagai sumber data yang jika diolah dengan baik dapat menghasilkan macam-macam informasi yang bermanfaat, yakni berupa data. Data tersebut dapat digunakan untuk penentuan keputusan, media untuk klarifikasi, dan bertukar pikiran bagi sesama pengguna Twitter yang biasanya menggunakan simbol # yang dibaca *hashtag* sebagai kata kunci. Belakangan ini, salah satu hal yang menjadi topik pembicaraan yang ramai dibahas pengguna Twitter adalah mengenai sektor transportasi umum.

Transportasi umum menjadi pilihan masyarakat di Jakarta yang padat penduduk. Menurut laporan Dinas Lingkungan Hidup Pemprov DKI Jakarta, Transjakarta merupakan salah satu transportasi umum yang memiliki jumlah penumpang harian terbanyak dibandingkan dengan transportasi umum lainnya (Annur, 2022). Transjakarta menjadi moda transportasi pilihan bagi penduduk kota Jakarta untuk melakukan mobilitas dalam kesehariannya. Hal tersebut membuat para pengguna Transjakarta dapat memberikan pendapat, kritik maupun saran dari pengalaman yang dialami oleh pengguna moda transportasi umum Transjakarta.

Penyebab dari banyaknya opini masyarakat terhadap transportasi Transjakarta adalah karena semakin banyaknya pengguna transportasi Transjakarta, yang diharapkan seimbang dengan ditingkatkannya pelayanan, infrastruktur, dan keamanan, sehingga opini negatif terhadap transportasi Transjakarta berkurang. Salah satu permasalahan yang terjadi pada transportasi Transjakarta yaitu beberapa halte mengalami penutupan sementara waktu dikarenakan ada proyek pembangunan rute MRT (*Mass Rapid Transit*). Selain hal itu, juga marak terjadi kasus asusila yang dialami oleh beberapa pengguna Transjakarta. Terkait permasalahan tersebut, banyak opini yang dikemukakan oleh masyarakat terhadap pelayanan transportasi umum Transjakarta. Penelitian ini bertujuan untuk menganalisis persepsi masyarakat khususnya pengguna media sosial Twitter melalui analisis sentimen dengan menggunakan *Metode Naïve Bayes* dan *Support Vector Machine (SVM)* yang berdampak untuk pelayanan Transjakarta.

Berdasarkan hasil implementasi dan pengujian menggunakan Metode *Naïve Bayes* yang telah dilakukan oleh Ikasari dkk. (2020) pada analisis sentimen, dengan jumlah tweet data latih sebanyak 90 tweet dan data uji sebanyak 10 tweet, didapatkan akurasi sistem sebesar 95.88%, sedangkan menurut Laurensz & Sediyono (2021), metode *Support Vector Machine (SVM)* merupakan model klasifikasi yang sangat populer karena menghasilkan akurasi yang baik. Proses utama dari Metode *SVM* adalah mencari batas yang memisahkan setiap kelas atau *hyperplane*. Hasil sentimen pada *Metode SVM* menggunakan kata kunci

"vaksinsinovac" mendapatkan persentase positif sebesar 96% dan negatif 4%, sedangkan persentase sentimen positif pada kata kunci "vaksinmerahputih" sebesar 98% dan negatif 2%. Dalam penelitian ini juga menunjukan bahwa hasil klasifikasi dengan menggunakan metode *Metode SVM* mendapatkan akurasi 84.41%.

Naïve Bayes merupakan sebuah model klasifikasi statistik yang dapat digunakan untuk memprediksi probabilitas keanggotaan suatu kelas. *Naïve Bayes* didasarkan pada teorema *bayes* yang memiliki kemampuan klasifikasi serupa dengan *Decision Tree* dan *Neural Network* (Dewi, 2019). Aturan Bayes didefinisikan sebagai aturan teoretis yang menghubungkan dua probabilitas bersyarat, $P\ (A|\ B)$ dan $P\ (B|\ A)$, satu sama lain. Ciri utama dari metode *Naïve Bayes* adalah asumsi yang kuat akan independensi dari masing-masing kondisi (Jo, 2019). Berikut formula perhitungan *Naïve Bayes*:

$$P\ (A|B) = \frac{P\ (B|A)\ P\ (A)}{P\ (B)} \quad (1)$$

Keterangan:

B	:	Data dengan *class* yang belum diketahui
A	:	Hipotesis data ialah sebagai *class* spesifik
$P\ (A \mid B)$	:	Probabilitas hipotesis A berdasar kondisi B (*Posterior Probability)*
$P\ (B \mid A)$	:	Probabilitas B berdasar kondisi pada hipotesis A (*Conditional Probaility*)
$P\ (A)$	:	Probabilitas hipotesis A (*Prior Probability*)
$P\ (B)$	:	Probabilitas B (*Evidence*)

Salah satu model dari *Naïve Bayes* yang sering digunakan dalam klasifikasi teks adalah *multinomial Naïve Bayes* (Sabrani & Bimantoro, 2020). *Multinomial Naïve Bayes* merupakan salah satu metode spesifik dari Metode *Naïve Bayes* yang menggunakan *conditional probability*, dengan menggunakan frekuensi kemunculan suatu kata pada suatu kelas (*raw term frequency*). Persamaan *Multinominal Naïve Bayes* dengan *conditional probability* menggunakan *add one* dan *Laplace smoothing* yaitu sebagai berikut (Fanesya dkk., 2019).

$$P\left(x_i|w_j\right) = \frac{N(x_i, w_j) + 1}{N(w_j) + |v|} \quad (2)$$

Keterangan:

$N(x_i, w_j) + 1$	:	Jumlah dari suatu kata query yang muncul dalam satu kelas, penambahan angka 1 digunakan untuk menghindari nilai 0.
Nw_j	:	Jumlah kata yang ada pada kelas w_j
$\|v\|$	:	Jumlah seluruh kata unik yang ada pada semua kelas

Formula *Multinomial Naïve Bayes* digunakan untuk melakukan pengujian klasifikasi, yakni sebagai berikut (Pratama dkk., 2024; Witten & Frank, 2002).

$$P\left(x_i|w_j\right) = P(x_i) \prod_{j=1}^{n} P\left(w_j|x_i\right) \quad (3)$$

Kemudian, penentuan kelasnya yaitu dengan memilih nilai maksimum, dengan formula:

$$V_{map} = \arg\max_{x_i \in X} P(x_i) \prod_{j=1}^{n} P\left(w_j|x_i\right) \quad (4)$$

Support Vector Machine (*SVM*) merupakan salah satu teknik dalam *machine learning* yang memiliki tingkat akurasi dan kualitas yang baik, sehingga sangat populer di antara algoritma lainnya (Laurensz & Eko Sediyono, 2021). Konsep *SVM* secara sederhana

dapat dijelaskan sebagai upaya untuk mencari *hyperplane* terbaik sebagai pemisah antara dua kelas pada *input space*. *SVM* mencoba mencari fungsi pemisah *(hyperplane)* dengan memaksimalkan jarak antar kelas. Dengan cara tersebut, *SVM* dapat menjamin generalisasi yang tinggi untuk data masa depan (Rokhman dkk., 2021). Proses analisis *SVM* akan dimulai dengan mengubah data teks yang ada menjadi data vektor, yang kemudian digabungkan untuk dilakukan pembobotan menggunakan *Term Frequency Inverse Document Frequency (TF-IDF)*. Keunggulan *SVM* dapat dilihat dari kemampuannya untuk mengidentifikasi *hyperplane* yang terpisah, sehingga dapat memaksimalkan margin kelas yang berbeda. SVM juga memiliki kelemahan yaitu masalah dengan karakteristik yang sama dapat mempengaruhi akurasi secara signifikan (Laurensz & Eko Sediyono, 2021). Namun ketika didapatkan data *non-linear*, *SVM* sulit mengklasifikasikan data. Dengan begitu penggunaan kernel dapat dijadikan solusi yang bertujuan untuk mentransformasikan data ke ruang berdimensi yang tinggi, dengan menjadikan data non linier terpisah secara *linier* (Awad & Khanna, 2015). Adapun tahapan dalam perhitungan Metode *SVM* dengan menggunakan fungsi kernel yaitu melalui beberapa langkah sebagai berikut (Manalu dkk., 2022).

1. Menentukan nilai α = 0.5, C = 1, λ = 0.5, *gamma* = 0.5 dan *epsilon* = 0.001
2. Menghitung matriks dengan Persamaan (5):

$$D_{ij} = y_i y_j + K(x_i, x_j) + \lambda^2 \quad (5)$$

dengan:

D_{ij} : Matriks data ke ij
y_i : Label data ke i
λ : lamda
$K(x_i, x_j)$: Fungsi Kernel

Adapun persamaan dalam menghitung fungsi kernel sebagai berikut (Awad & Khanna, 2015).

$$K(x_i, x_j) = \sum_{i=1}^{n} x_i{}^T x_j \quad (6)$$

3. Pada Persamaan (6) sebagai perhitungan nilai *error*:

$$E_i = \sum_{i=1}^{n} \alpha_i D_{ij} \quad (7)$$

4. Perhitungan untuk nilai delta alpha dengan Persamaan (8):

$$\delta\alpha_i = \min\{\max[\gamma(1 - E_i), -\alpha_i], C - \alpha_i\} \quad (8)$$

Keterangan:

E_i : Nilai *error* pada data ke i
γ : Tingkat pembelajaran
$max(i)\ D_{ij}$: Nilai maksimum matriks hessian

5. Menghitung nilai alpha baru dengan Persamaan (9):

$$\alpha_i = \alpha_i + \delta\alpha_i \quad (9)$$

6. Persamaan (10) untuk mencari nilai bias (b):

$$b = -\frac{1}{2}(W * x^+ + W * x^-) \quad (10)$$

7. Pengujian terhadap data
8. Perhitungan keputusan yakni dengan menggunakan Persamaan (11) dapat ditentukan sebagai kelas positif jika perolehan $h(x) \geq 0$ maka nilai $sign\ h(x)$ adalah 1, sedangkan jika hasil perhitungan keputusan memiliki nilai lebih kecil dari 0 maka $sign\ h(x)$ adalah -1.

$$h(x) = \sum_{i}^{n} \alpha_i \; y_i \, K(x, x_i) + b \tag{11}$$

Term Frequency-Inverse Document Frequency (TF-IDF) merupakan suatu metode untuk menghitung bobot dari kata yang digunakan untuk menghitung nilai setiap kata. *Term Frequency (TF) dan Inverse Document Frequency (IDF)* memiliki suatu perbedaan yaitu sebagai berikut (Haj & Amrizal, 2020).

1. *Term Frequency (TF):* Semakin bertambah banyak jumlah kemunculan kata, maka semakin besar kata-kata itu akan berpengaruh pada dokumen, dan begitu juga sebaliknya.
2. *Inverse Document Frequency (IDF):* Semakin banyak dokumen yang mengandung suatu kata tertentu, semakin kecil pengaruh kata tersebut pada dokumen, dan begitu sebaliknya.

TF-IDF (*Term Frequency-Inverse Document Frequency*) dihitung menggunakan rumus sebagai berikut (Febriyani & Februariyanti, 2023).

$$W(d,t) = TF(d,t) \times IDFf(t) \tag{12}$$

dengan:

$W(d,t)$: Bobot dokumen ke-d terhadap kata ke-t
$TF(d,t)$: Teks frekuensi
$IDFf(t)$: Teks frekuensi dalam dokumen

Lexicon Based adalah sekumpulan kata yang biasa digunakan untuk mengekspresikan emosi positif atau sentimen, dengan nilai yang ditetapkan untuk setiap kata. *Lexicon based* merupakan kamus data yang berisi kumpulan kata bahasa Indonesia yang diberi nilai berupa angka yang berkaitan dengan emosi positif atau negatif. Keuntungan dari pendekatan ini adalah pelabelan kalimat dapat dilakukan secara otomatis sehingga kumpulan data yang banyak jumlahnya lebih cepat diproses. Kamus berisi sekumpulan kata dengan bobot yang ditetapkan untuk setiap kata (Sharda dkk., 2018).

Grid Search Cross Validation merupakan langkah dalam pencarian kombinasi parameter optimal dalam pemodelan analisis sentimen algoritma pembelajaran mesin (Friedman, 2009). Proses *Grid Search Cross Validation* melibatkan pemisahan data set menjadi subset untuk validasi silang (*cross-validation*). Setiap kombinasi parameter yang berbeda diuji pada sekumpulan data latih dengan metrik evaluasi yang sesuai. Dengan mencoba berbagai kombinasi parameter, pendekatan ini membantu menghindari pemilihan parameter yang kurang tepat dan dapat meningkatkan performa model ketika memprediksi sentimen menjadi lebih baik (Géron, 2019). Sebuah metode biasa digunakan untuk melakukan perhitungan presisi pada konsep data *mining*. Rumus metode ini dihitung dengan 4 *output* yaitu: *recall*, presisi, akurasi dan tingkat kesalahan evaluasi model klasifikasi didasarkan pada tes untuk memperkirakan objek yang benar dan salah (Dewi, 2019).

Tabel 1. *Confusion Matrix*

Hasil Prediksi	**Nilai Aktual**	
	Positif	**Negatif**
Positif	TP	FP
Negatif	FN	TN

Sumber:(Waluyo & Prihandoko, 2017)

Adapun formula untuk menghitung akurasi, presisi, *recall* dan *f1-Score* sebagai berikut:

$$\text{Akurasi} = \frac{TP + TN}{TP + FP + FN + TN} \tag{13}$$

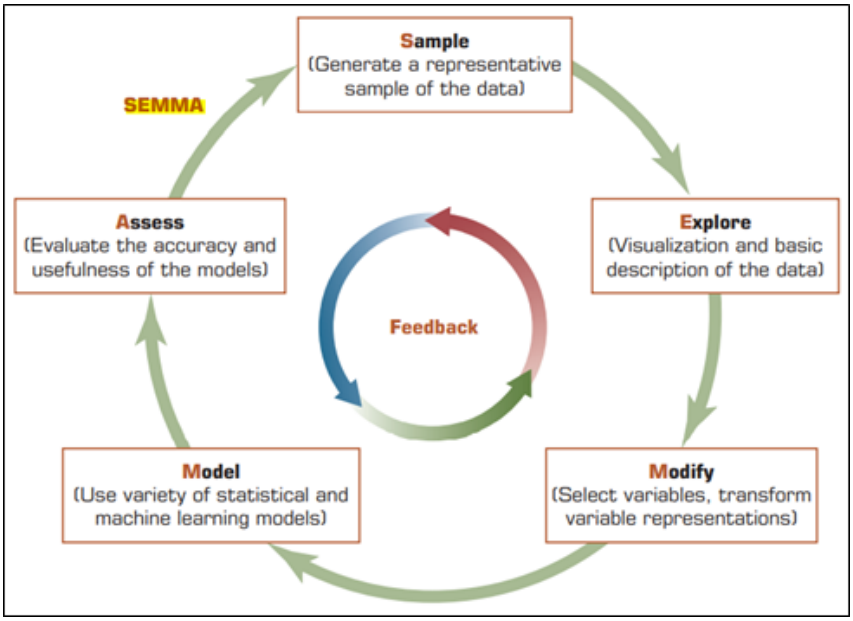
$$\text{Presisi} = \frac{TP}{TP + FP} \quad (14)$$

$$Recall = \frac{TP}{TP + FN} \quad (15)$$

$$f1 - score = 2 \times \frac{presisi \times recall}{presisi + recall} \quad (16)$$

METODE PENELITIAN

Metode yang digunakan dalam penelitian ini mengacu pada *Data Mining Process* (SEMMA). SEMMA adalah suatu metode data mining yang dikembangkan oleh SAS instituate. Kepanjangan dari SEMMA ialah *Sample, Explore, Modify, Model dan Assess*. SEMMA juga dapat membantu untuk memberikan solusi terhadap masalah dan tujuan bisnis (Antonio dkk., 2022). Tahapan yang dilakukan pada penelitian ini dapat dijelaskan pada Gambar 1.



Gambar 1. Proses SEMMA Data Mining
(Sharda dkk., 2018)

Berikut tahapan hasil penelitian dengan menerapkan Metode SEMMA.

1. *Sample*

   Dataset diperoleh dengan cara *crawling* data pada media sosial Twitter. Kemudian melakukan proses *crawling* data menggunakan *library Snscrape* pada bahasa pemograman *Python* dengan *Jupyter Notebook*. Pencarian data menuliskan kata kunci "transjakarta" dengan rentang waktu pada bulan Maret-April. Setelah mendapatkan hasil data *tweet* dalam proses *crawling* data, maka simpan data tersebut dalam file yang berformat '.csv'.

2. *Explore*

   Pada tahapan ini data Twitter berisi 15 kolom yang berisi rincian dari sebuah *tweet*. Pada penelitian ini hanya digunakan 2 kolom yaitu 'Date' yang merupakan waktu pengguna memposting *tweet* dan 'Tweet' yang merupakan isi konten *tweet* pengguna yang di posting pada media sosial Twitter. Berikut adalah hasil dari filter kolom yang dilakukan.

	Date	Tweet
0	2023-03-30 23:54:42+00:00	@treasuremenfess kl perkiraan total ini tergan...
1	2023-03-30 23:53:37+00:00	@audnmonika Pagi kak. Terima kasih atas inform...
2	2023-03-30 23:52:54+00:00	@andiifaauzan Hai Kak. Pembelian kartu Jak Lin...
3	2023-03-30 23:52:31+00:00	@PT_Transjakarta tolonglah min bus tipe PPD ad...
4	2023-03-30 23:47:49+00:00	@PT_Transjakarta Selamat pagi, kalo mau beli k...

Gambar 2. Hasil Pengambilan Dua Kolom Dataset

3. *Modify*
   Pada tahap ini dilakukan modifikasi berupa *text preprocessing* pada data set dengan tujuan agar data set menjadi terstruktur dan lebih dikenali bentuknya oleh sistem komputer untuk diolah lebih lanjut. Modifikasi data tweet yang belum diolah akan melalui beberapa proses preprocessing diantaranya sebgai berikut :
   a. *Case Folding:* Proses ini dilakukan untuk mengganti karakter huruf pada *dataset* menjadi huruf kecil, seperti pada Tabel 2

Tabel 2. Hasil Proses *Case Folding*

Sebelum	Sesudah
Tolonglah @PT_Transjakarta klo buat peraturan tuh yg tegas. Klo harus antre ya semua harus antre dong. Jgn ada yg dibiarkan lewat krn berani melawan. Sakit hati banget liat kita cape antre tp yg nyela didiemin aja. Kejadian di Halte Apt Kalibata Jumat (31/3) jam 06.16.	tolonglah @pt_transjakarta klo buat peraturan tuh yg tegas. klo harus antre ya semua harus antre dong. jgn ada yg dibiarkan lewat krn berani melawan. sakit hati banget liat kita cape antre tp yg nyela didiemin aja. kejadian di halte apt kalibata jumat (31/3) jam 06.16.

   b. *Cleaning*: Proses ini dilakukan pembersihan data set seperti menghapus *retweet* (RT), *mention*, *link*, *hastag*, url, tanda baca dan *whitespace*, hasil *cleaning* dapat dilihat pada Tabel 3.

Tabel 3. Hasil Proses *Cleaning*

Sebelum	Sesudah
tolonglah @pt_transjakarta klo buat peraturan tuh yg tegas. klo harus antre ya semua harus antre dong. jgn ada yg dibiarkan lewat krn berani melawan. sakit hati banget liat kita cape antre tp yg nyela didiemin aja. kejadian di halte apt kalibata jumat (31/3) jam 06.16.	tolonglah transjakarta klo buat peraturan tuh yg tegas klo harus antre ya semua harus antre dong jgn ada yg dibiarkan lewat krn berani melawan sakit hati banget liat kita cape antre tp yg nyela didiemin aja kejadian di halte apt kalibata jumat jam

   c. *Tokenize*: Proses ini untuk memecah data set menjadi potongan kata yang bertujuan untuk mempermudah proses selanjutnya.
   d. *Normalize*: Proses ini mengubah kata slang dan singkatan menjadi kata baku yang sesuai agar mempunyai arti yang seragam dengan menggunakan kamus NLP Bahasa

Indonesia yang berbentuk *file* '.xlsx'. Tabel 4 merupakan proses melakukan normalisasi.

Tabel 4. Hasil Proses *Normalize*

Sebelum	Sesudah
[tolonglah, transjakarta, klo, buat, peraturan, tuh, yg, tegas, klo, harus, antre, ya, semua, harus, antre, dong, jgn, ada, yg, dibiarkan, lewat, krn, berani, melawan, sakit, hati, banget, liat, kita, cape, antre, tp, yg, nyela, didiemin, aja, kejadian, di, halte, apt, kalibata, jumat, jam]	[tolonglah, transjakarta, kalau, buat, peraturan, tuh, yang, tegas, kalau, harus, antre, iya, semua, harus, antre, dong, jangan, ada, yang, dibiarkan, lewat, karena, berani, melawan, sakit, hati, banget, liat, kita, cape, antre, tetapi, yang, nyela, diamkan, saja, kejadian, di, halte, apartemen, kalibata, jumat, jam]

e. *Stopworod Removeal*: Proses ini untuk menghapus kata-kata yang sering muncul tetapi tidak memiliki arti penting dan maknanya tidak berpengaruh pada sistem, seperti „di", „oh", „pada", dan lain-lain.
f. *Stemming*: Proses ini digunakan untuk mengembalikan kata ke dalam bentuk dasar dari kata tersebut. Seperti kata imbuhan "ter", "men", "ber", dan lain-lain, dapat dilihat pada Tabel 5.

Tabel 5. Hasil Proses *Stemming*

Sebelum	Sesudah
[tolonglah, transjakarta, peraturan, tuh, antre, iya, antre, dibiarkan, berani, melawan, sakit, hati, banget, liat, cape, antre, nyela, diamkan, kejadian, halte, apartemen, kalibata, jumat, jam]	[tolong, transjakarta, atur, tuh, antre, iya, antre, biar, berani, lawan, sakit, hati, banget, liat, cape, antre, nyela, diam, jadi, halte, apartemen, kalibata, jumat, jam]

Tahapan *modify* menghasilkan total 5366 tweet dengan menggunakan pemograman Python, kemudian setelah dilakukan *text preprocessing* menjadi 5049 tweet pada dataset.

4. *Model*
   Klasifikasi data bersih menurut kelasnya untuk menentukan apakah termasuk opini positif atau negatif. Dalam proses tersebut akan menggunakan kamus yaitu *lexicon based*. setelah dataset bersih diberi antara positif dan negatif, maka selanjutnya akan divisualisasikan ke bentuk *wordcloud*. Kemudian dataset yang sudah diberikan label akan dilakukan proses pembagian data menjadi data latih dan data uji untuk dilakukan pemodelan dengan *Algoritma Naïve Bayes* dan *SVM*.
5. *Asses*
   Evaluasi pemodelan yang ada dengan cara membandingkan hasil dari prediksi terhadap data uji menggunakan label sentimen yang telah didapatkan sebelumnya. Hasilnya berupa nilai confusion matrix yaitu akurasi, presisi, *recall* dan *f1-score*.

PEMBAHASAN DAN HASIL

Tahapan berikutnya adalah melakukan pelabelan atau pemberian kelas pada dataset mengunakan *lexicon based* dengan kamus Lexicon Inset yang bersumber dari Koto & Rahmaningtyas (2017) dan menghasilkan sentiment positif sejumlah 2228 dan negatif 2821.

Implementasi Metode *SVM* dan *Naïve Bayes*

Perhitungan *TF-IDF* untuk *SVM* dilakukan dengan nilai $\alpha = 0.5, C = 1, \lambda = 0.5, \gamma = 0.5\ dan\ \varepsilon = 0.001$ pada 3 dokumen seperti di Tabel 6.

Tabel 6. Pembobotan Kata TF-IDF Perhitungan Manual SVM

	Tweet	y	Total Dokumen
D1	pagi terima kasih informasinya	1	3
D2	transjakarta tolong minimal bus tipe ppd ada petugas	-1	
D3	beli kartu jaklingko divending machine terima kasih	?	

Term	TF			DF	N/DF	IDF	(w) = TF*IDF		
	D1	D2	D3				D1	D2	D3
pagi	1	0	0	1	3	0,477	0,477	0,000	0,000
terima	1	0	1	2	1,5	0,176	0,176	0,000	0,176
kasih	1	0	1	2	1,5	0,176	0,176	0,000	0,176
informasinya	1	0	0	1	3	0,477	0,477	0,000	0,000
transjakarta	0	1	0	1	3	0,477	0,000	0,477	0,000
tolong	0	1	0	1	3	0,477	0,000	0,477	0,000
minimal	0	1	0	1	3	0,477	0,000	0,477	0,000
bus	0	1	0	1	3	0,477	0,000	0,477	0,000
tipe	0	1	0	1	3	0,477	0,000	0,477	0,000
Ppd	0	1	0	1	3	0,477	0,000	0,477	0,000
Ada	0	1	0	1	3	0,477	0,000	0,477	0,000
petugas	0	1	0	1	3	0,477	0,000	0,477	0,000
Beli	0	0	1	1	3	0,477	0,000	0,000	0,477
kartu	0	0	1	1	3	0,477	0,000	0,000	0,477
jaklingko	0	0	1	1	3	0,477	0,000	0,000	0,477
divending	0	0	1	1	3	0,477	0,000	0,000	0,477
machine	0	0	1	1	3	0,477	0,000	0,000	0,477

Setelah itu, tahapan yang dilakukan yaitu proses klasifikasi SVM dengan fungsi kernel, yang dimana setiap data akan dihitung terhadap data itu sendiri serta antara data lainnya dengan menggunakan Persamaan (6). Berikut Tabel 7 dari fungsi kernel.

Tabel 7. Fungsi Kernel

	D1	D2
D1	$K(D1, D1) = 0.320$	$K(D1, D2) = 0$
D2	$K(D2, D1) = 0$	$K(D2, D2) = 1.820$

Langkah selanjutnya setelah mendapatkan hasil perhitungan fungsi kernel kemudian melakukan proses perhitungan matriks dengan Persamaan (5) sebagai berikut

$$(D)_{D1,D1} = (1)(1) + (0.320) + 0.5^2 = 0.570$$

Dengan cara serupa, data lainnya dihitung dengan hasil yang tercantum pada Tabel 8.

Tabel 8. Hasil Perhitungan Matriks

	D1	D2
D1	0.570	0.250
D2	0.250	2.070

Setelah mendapatkan hasil perhitungan nilai matriks, yang dilakukan adalah

mencari nilai *error* dengan menggunakan Persamaan (8) sebagai berikut:

$$E_{D1} = (1)(0.5)(0.820) = 0.410 \text{ dan } E_{D2} = 1.160$$

Tahapan selanjutnya yaitu mencari nilai dari *delta alpha* dengan menggunakan Persamaan (8) sebagai berikut:

$$\begin{aligned}\delta\alpha_{D1} &= min\{max[0.5(1-0.410), -0.5], 1-0.5\} \\ &= min\{max[0.295, -0.5], 0.5\} \\ &= \min[0.295 \times 0.5] \\ &= 0.295\end{aligned}$$

serta $\delta\alpha_{D2} = -0.08$. Kemudian, perhitungan nilai *alpha* baru dilakukan dengan Persamaan (9) sebagai berikut:

$$\alpha'_{D1} = 0.5 + 0.295 = 0.795 \text{ dan } \alpha'_{D2} = 0.420.$$

Nilai bias dapat dicari menggunakan Persamaan (10) dengan terlebih dahulu menentukan nilai W yang dimana W^+ adalah bobot *dot product* dengan nilai *alpha* terbesar pada kelas positif dan W^- adalah bobot *dot product* dengan nilai *alpha* terbesar dari kelas negatif. Lalu perhitungannya yaitu sebagai berikut:

$$W * x^+ = (0.795 \times 1 \times 0.320) + (0.420 \times (-1) \times 0) = 0.255$$

$$W * x^- = (0.795 \times 1 \times 0) + (0.420 \times (-1) \times 1.820) = -0.764$$

$$b = -\frac{1}{2}\big(0.255 + (-0.764)\big) = 0.255$$

Tahapan selanjutnya melakukan pengujian data pada data sampel uji pada kalimat "beli kartu jaklingko divending machine terima kasih". Pada saat kalimat dipecah menjadi perkata dan dimasukkan dalam bentuk TF-IDF yang dapat dilihat pada Tabel 9 sebagai berikut.

Tabel 9. Sampel Data Uji

D	TF-IDF																	y
	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13	T14	T15	T16	T17	
Dx	0	0.176	0.176	0	0	0.	0	0	0	0	0	0	0.477	0.477	0.477	0.477	0.477	?

Pada proses awal tahap menguji adalah menghitung *dot product* data uji dengan semua data latih menggunakan fungsi kernel sebagai berikut:

$$\begin{aligned}K(D1, W_{D1}) = &(0 \times 0.477) + (0.176 \times 0.176) + (0.176 \times 0.176) + (0 \times 0.176) \\ &+ (0 \times 0) + (0 \times 0) + (0 \times 0) + (0 \times 0) + (0 \times 0) \\ &+ (0 \times 0) + (0 \times 0) + (0.477 \times 0) + (0.477 \times 0) + (0.477 \times 0) \\ &+ (0.477 \times 0) + (0.477 \times 0) = 0.062\end{aligned}$$

dan $K(D2, W_{D1}) = 0$. Tahapan terakhir yaitu perhitungan fungsi keputusan dengan Persamaan (11) sebagai berikut:

$$h(x) = sign((0.795 \times 1 \times 0.062) + 0.255 + (0.420 \times (-1) \times 0) + 0.255) = 1$$

Berdasarkan dari perhitungan fungsi keputusan, maka data uji tersebut didapatkan hasilnya kedalam kelas 1 atau positif.

Metode selanjutnya adalah *Naïve Bayes* yakni dengan melakukan percobaan perhitungan dengan menggunakan beberapa sampel data dari *dataset* yang terdapat penelitian ini. Berikut Tabel 10 merupakan sampel data untuk dilakukan perhitungan manual *Naïve Bayes*.

Tabel 10. Sampel Data Manual *Naïve Bayes*

	Dokumen	Tweet	Label
Data Latih	1	pagi kak terima kasih informasinya	Positif
	2	kak beli kartu jaklingko beli di vending machine	Positif
	3	terima kasih transjakarta tolong minimal bus tipe ppd ada petugasnya	Negatif
Data Uji	4	Halo admin tolong benerin acnya	?

Setelah itu, beberapa percobaan rasio dilakukan dengan 70% data latih dan 30% data uji, 80% data latih dan 20% data uji, 90% data latih dan 10% data uji. Berikut Tabel perbandingan data latih dan data uji yang digunakan dengan menggunakan algoritma *Naïve Bayes Classifier* dan SVM.

Tabel 11. Perbandingan Data Latih dan Data Uji *Naïve Bayes* dan SVM

Data Latih	Data Uji
70%	30%
80%	20%
90%	10%

Tahap pertama yaitu menghitung probabilitas sentimen positif dan negatif, diperoleh:
Probabilitas kelas positif:

$$P(Positif) = \frac{2}{3}$$

Probabilitas kelas negatif:

$$P(Negatif) = \frac{1}{3}$$

Setelah mendapatkan *prior* atau probabilitas kemudian menghitung *Likelihood* atau *conditional probabilities* menggunakan Persamaan (3) pada data uji yang akan dihitung, seperti pada Tabel 13.

Tabel 13. Nilai *Conditional Probabilities*

Kata	Conditional Probabilities	
	Positif	**Negatif**
halo	$\frac{1}{30}$	$\frac{1}{27}$
admin	$\frac{1}{30}$	$\frac{1}{27}$
tolong	$\frac{1}{30}$	$\frac{2}{27}$
benerin	$\frac{1}{30}$	$\frac{1}{27}$
acnya	$\frac{1}{30}$	$\frac{1}{27}$

Nilai tersebut berguna untuk menghitung data uji untuk dikategorikan di kelas positif atau

negatif dengan menggunakan Persamaan (4) sebagai berikut:

$$V_{map} = \arg\max_{x_i \in X} \left\{ P(Positif|dokumen\ 4) = \left(\frac{1}{30^5}\right), P(Positif|dokumen\ 4 = \left(\frac{2}{27^5}\right) \right\}$$
$$= \left(\frac{2}{27^5}\right)$$

Berdasarkan hasil dari perhitungan terhadap data uji untuk menentukan kalimat tersebut termasuk positif atau negatif dengan menggunakan nilai yang lebih besar yaitu nilai $\frac{2}{27^5}$, maka sentimen dari data uji tersebut termasuk pada label "Negatif".

Evaluasi Model

Pada tahapan *assess* akan dilakukan evaluasi terhadap model penelitian. Hasil evaluasi dari kedua metode antara *Naïve Bayes* dan *Support Vector Machine* akan menggunakan *Confusion Matrix* yang memuat nilai akurasi, presisi, *recall*, dan *f1-score* dari data uji. Hasil dari Metode *Naïve Bayes* pada rasio 70%:30% menjadi rasio yang terbaik dengan mendapatkan akurasi 76.43%, sedangkan Metode SVM dengan rasio 80%:20% menghasilkan akurasi 84.95%. Tabel 14 menyajikan perbandingan hasil dari beberapa rasio data set terdiri dari akurasi, presisi, *recall* dan *f1-score*, yang dihitung menggunakan Persamaan (13)-(16).

Tabel 14. Perbandingan Hasil Evaluasi *Naïve Bayes* dengan *SVM*

Metode	Data Latih:Data Uji	Akurasi	Presisi	Recall	F-1 Score
Naïve Bayes	70% : 30%	76.43%	78%	68%	73%
	80% : 20%	76.13%	77%	68%	72%
	90% : 10%	75.04%	73%	70%	72%
SVM	70% : 30%	84.35%	84.17%	81.40%	82.76%
	80% : 20%	84.95%	83%	83%	83%
	90% : 10%	82.97%	80.43%	81.85%	81.14%

Dari hasil beberapa skenario pada rasio data latih dan data uji yang dilakukan, maka didapatkan akurasi terbaik yaitu pada rasio 70%:30% dengan begitu *Confusion Matrix* dari Metode *Naïve Bayes* dan *SVM* pada rasio 70%:30% dan 80%:20% mendapatkan hasil pada Tabel 15.

Tabel 15. *Confusion Matrix Naive Bayes* dengan *SVM*

Metode	Nilai Prediksi	Nilai Aktual	
		Positif	**Negatif**
Naïve Bayes	**Positif**	478 (TP)	136 (FP)
	Negatif	221(FN)	680 (TN)
SVM	**Positif**	380 (TP)	76 (FP)
	Negatif	76 (FN)	478 (TN)

KESIMPULAN DAN SARAN

Data bersih yang telah diberikan label berdasarkan *Lexicon Based* menghasilkan 2228 tweet kelas positif dan 2821 tweet kelas negatif. Metode *Support Vector Machine* mendapatkan hasil akurasi 84.95%, presisi 83%, *recall* 83% dan *f1-score* 83%, sedangkan Metode *Naïve Bayes* mendapatkan hasil akurasi 76.43%, presisi 78%, *recall* 68% dan *f1-score* 73%. Dengan demikian, metode yang lebih unggul adalah *Support Vector Machine*. Penelitian ini terbatas hanya menggunakan metode *Support Vector Machine (SVM*) dan *Naïve Bayes*. Untuk penelitian selanjutnya, disaranakan untuk menggunakan metode yang lain seperti *Random Forest, K-Nearest Neighbor (KNN),* dan lain sebagainya. *Random*

Forest merupakan salah satu algoritma terbaik untuk klasifikasi data yang besar atau memiliki banyak fitur dengan akurasi tinggi, sedangkan metode KNN untuk pengkategorisasian teks dan algoritma dapat digunakan untuk mempelajari struktur data yang ada serta mengkategorikan dirinya sendiri.

DAFTAR RUJUKAN

Annur, C. M. (2022, November). *Ini rute bus transjakarta dengan jumlah penumpang terbanyak pada 2021*. databoks.katadata.co.id.

Annur, C. M. (2023). *Pengguna twitter di indonesia capai 24 juta hingga awal 2023, peringkat berapa di dunia?* databoks.katadata.co.id.

Antonio, N., de Almeida, A., & Nunes, L. (2022). Data mining and predictive analytics for e-tourism. *In Handbook of E-Tourism*, 531–555. https://doi.org/https://doi.org/10.1007/978-3-030-48652-5_29

Awad, M., & Khanna, R. (2015). *Efficient learning machines: theories, concepts, and applications for engineers and system designers*. Springer nature.

Dewi, S. (2019). Komparasi metode algoritma data mining pada prediksi uji kelayakan credit approval pada calon nasabah kredit perbankan. *Jurnal Khatulistiwa Informatika*, *7*(1). https://doi.org/10.31294/jki.v7i1.5744

Febriyani, E., & Februariyanti, H. (2023). Analisis sentimen terhadap program kampus merdeka menggunakan naive bayes classifier di twitter. *Jurnal TeknoKompak*, *17*(1), 25–38. https://doi.org/10.30865/json.v4i2.5381

Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction. In *Springer*.

Géron, A. (2019). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow*. " O'Reilly Media, Inc."

Haj, A. S. A., & Amrizal, V. (2020). Analisis sentimen kinerja kpu pemilu 2019 menggunakan algoritma k-means dengan algoritma confix stripping stemmer. *Journal of Innovation Information Technology and Application (JINITA)*, *2*(1), 9–18.

Ikasari, D., Fajarwati, Y., & Widiastuti. (2020). Analisis sentimen dan klasifikasi tweets berbahasa indonesia terhadap transportasi umum mrt jakarta menggunakan naïve bayes classifier. *Jurnal Ilmiah Informatika Komputer*, *25*(1), 64–75. https://doi.org/10.35760/ik.2020.v25i1.2427

Jo, T. (2019). *Text mining: Concepts, implementation, and big data challenge* (Vol. 45).

Koto, F., & Rahmaningtyas, G. Y. (2017). Inset lexicon: Evaluation of a word list for Indonesian sentiment analysis in microblogs. *2017 International Conference on Asian Language Processing (IALP)*, 391–394.

Laurensz, B., & Sediyono, E. (2021). Analisis sentimen masyarakat terhadap tindakan vaksinasi dalam upaya mengatasi pandemi covid-19. *Jurnal Nasional Teknik Elektro dan Teknologi Informasi*, *10*(2), 118–123. https://doi.org/10.22146/jnteti.v10i2.1421

Manalu, D. R., Tobing, M. C. L., & Yohanna, M. (2022). Analisis sentimen twitter terhadap wacana penundaan pemilu dengan metode support vector machine. *METHOMIKA: Jurnal Manajemen Informatika & Komputerisasi Akuntansi*, *6*(2), 149–156.

Pratama, D. R. S., Munandar, T. A., & Ramdhania, K. F. (2024). Multinomial naive bayes algorithm for indonesian language sentiment classification related to jakarta international stadium (jis). *International Journal of Information Technology and Computer Science Applications*, *2*(1), 12–22.

Rokhman, K. A., Berlilana, B., & Arsi, P. (2021). Perbandingan metode support vector

machine dan decision tree untuk analisis sentimen review komentar pada aplikasi transportasi online. *Journal of Information System Management (JOISM)*, *2*(2), 1–7.

Sabrani, A., & Bimantoro, F. (2020). Multinomial naïve bayes untuk klasifikasi artikel online tentang gempa di indonesia. *Jurnal Teknologi Informasi, Komputer, dan Aplikasinya (JTIKA)*, *2*(1), 89–100.

Sharda, R., Delen, D., & Turban, E. (2018). *Business intelligence, analytics, and data science: a managerial perspective*. pearson.

Witten, I. H., & Frank, E. (2002). Data mining: practical machine learning tools and techniques with java implementations. *Acm Sigmod Record*, *31*(1), 76–77.