Jurnal Statistika dan Aplikasinya Volume 6 Issue 2, December 2022 e-ISSN: 2620-8369

Received: 30 October 2022
Revised: 19 December 2022
Accepted: 29 December 2022
Published: 31 December 2022

Analisis Sentimen pada Program Transportasi Publik JakLingko dengan Metode *Support Vector Machine*

Faroh Ladayya[1, a)], Dania Siregar[1, b)], Wiligis Eka Pranoto[1, c)], Hilmy Dzaky Muchtar[1, d)]

[1]*Program Studi Statistika Fakultas Ilmu Pengetahuan Alam Universitas Negeri Jakarta*
Jl. Rawamangun Muka, Kota Jakarta Timur, DKI Jakarta, 13220

E-mail: [a)] farohladayya@unj.ac.id, [b)]dania-siregar@unj.ac.id, [c)]wiligisekapranoto_1314620022@mhs.unj.ac.id, [d)]hilmydzakymuchtar_1314620017@mhs.unj.ac.id

Abstract

As a metropolitan city with high mobility, public transportation plays an important role in facilitating economic, business and government activities in DKI Jakarta. DKI Jakarta provincial government launched the JakLingko program to create an integrated, convenient, efficient, and affordable public transportation system. Knowledge of public opinion can help improve the service quality of the JakLingko program. The use of social media is becoming very popular nowadays. Through social media, anyone can easily express their opinion about an issue. It is used to obtain objective and latest public opinion. Sentiment analysis is a method that can be used to analyze public opinion. Through sentiment analysis whose data was collected from *Twitter*, it can be seen how the public opinion toward JakLingko program. In this study, public sentiment will be classified into positive sentiment or negative sentiment. As for the classification, the Support Vector Machine (SVM) algorithm is used. Through the 5-fold-cross-validation process, the accuracy score of the training data is obtained to be 0,711 and the accuracy score is relatively stable for every fold. The accuracy of the training data is also close, which is 0,703. This score can be categorized as a good score. The topic that is widely discussed and has positive sentiment is the free-charged Jaklingko public transportation that heavily eased people. Meanwhile, the topic that is widely discussed and has negative sentiment is people's complaints about the use of Jaklingko cards and the lowly service of public transportation.

Keywords: JakLingko, Sentiment Analysis, Support Vector Machine

Abstrak

Sebagai kota metropolitan dengan mobilitas tinggi, transportasi umum berperan penting dalam memperlancar kegiatan ekonomi, bisnis, dan pemerintahan di DKI Jakarta. Pemerintah Provinsi DKI Jakarta mencanangkan program JakLingko untuk menciptakan sistem transportasi publik yang terintegrasi, nyaman, efisien, dan terjangkau. Pengetahuan tentang opini publik dapat membantu meningkatkan kualitas layanan program JakLingko. Penggunaan media sosial menjadi sangat populer saat ini. Melalui media sosial, siapa pun dapat dengan mudah

mengungkapkan pendapatnya tentang suatu masalah. Hal ini digunakan untuk memperoleh opini publik yang objektif dan terkini. Analisis sentimen merupakan metode yang dapat digunakan untuk menganalisis opini publik. Melalui analisis sentimen yang datanya dikumpulkan dari *Twitter*, dapat diketahui bagaimana opini publik terhadap program Jaklingko. Melalui proses klasifikasi data *tweet* JakLingko dengan 5-*fold cross-validation*, didapatkan rata-rata akurasi pada data *training* sebesar 0,711 dan nilai akurasinya stabil pada setiap *fold*. Akurasi yang diterapkan pada data *training* juga mendapatkan nilai yang tidak jauh berbeda yaitu 0,703. Nilai tersebut dapat dikategorikan sebagai nilai yang baik. Topik yang banyak dibicarakan dan memiliki sentimen positif adalah tarif angkot jaklingko gratis yang sangat memudahkan masyarakat. Sedangkan topik yang banyak dibicarakan pada sentimen negatif keluhan tentang penggunaan kartu JakLingko serta angkot yang masih belum baik pelayanannya.

Kata-kata kunci: JakLingko, Sentiment Analysis, Support Vector Machine.

PENDAHULUAN

Transportasi menjadi permasalahan yang sedang dialami oleh negara berkembang, seperti Indonesia. Sebagai ibukota Indonesia, Jakarta termasuk kota yang padat penduduk. Diiringi dengan semakin tingginya pemakaian kendaraan pribadi yang menyebabkan terjadinya kemacetan. Transportasi publik merupakan salah satu solusi dari permasalahan transportasi yang dibutuhkan untuk aktivitas di kota besar seperti Jakarta. Penggunaan transportasi publik dapat menurunkan tingkat kemacetan, mengurangi kadar polusi, dan pengurangan penggunaan bahan bakar (Triana, 2022). Namun masyarakat di Jakarta masih enggan menggunakan transportasi untuk menunjang kegiatan sehari-hari. Berdasarkan Jakarta Open Data, persentase dari penggunaan transportasi publik di Jakarta hanya berkisar sekitar 2% dari keseluruhan perjalanan yang ada di DKI Jakarta. Terlebih saat pandemi COVID-19, penggunaan transportasi mengalami penurunan yang signifikan. Pada bulan April 2020 penumpang MRT turun sebesar 94%, penumpang KRL turun hingga 78,69%, dan penumpang LRT turun hingga 93,5% (Simanjuntak, 2022). Untuk meningkatkan penggunaan transportasi publik diperlukan perbaikan pelayanan. Dalam meningkatkan pelayanan transportasi publik di Jakarta, Pemerintah Provinsi DKI Jakarta melakukan integrasi dengan PT Transport Jakarta, PT MRT Jakarta (Perseroda), PT Jakarta Propetindo, PT Moda Kemerdekaan Jabodetabel Transportation, dan PT Kereta Api Indonesia (Persero) bekerjasama menjadi PT JakLingko Indonesia yang mengintegrasikan tarif dan pembayaran elektronik antara moda transportasi MRT, Transjakarta (TJ), LRT Jakarta, KAI Commuter, Railink Soekarno-Hatta, dan Mikrotrans atau angkot. Masyarakat hanya perlu menggunakan satu skema pembayaran untuk menaiki berbagai angkutan ibu kota. Harapannya melalui program JakLingko, minat warga DKI Jakarta untuk menggunakan transportasi umum akan meningkat. Hal ini selaras dengan tugas pemerintah daerah untuk memberikan pelayanan yang lebih baik kepada masyarakat secara adil, merata, cepat dan tepat, serta menjalin kerjasama dengan pihak swasta (Unit Pelayanan Statistik, 2019). Dalam upaya peningkatan pelayanan JakLingko, pemerintah provinsi DKI Jakarta telah melakukan survey tentang kepuasan warga DKI Jakarta terhadap program JakLingko pada tahun 2019. Selain melalui survey secara langsung, penggunaan media sosial dapat menjadi pilihan untuk menjaring opini publik yang lebih efisien, objektif dan *realtime*. Masyarakat saat ini banyak memanfaatkan media sosial untuk saling berinteraksi, mengeluarkan pendapat dan respon terhadap berbagai topik. Berdasarkan data *Digital, Social & Mobile*, pengguna media sosial di Indonesia mencapai 150 juta pengguna dimana lebih dari 52% diantaranya adalah pengguna *twitter* (DataReportal, 2022). Potensi media sosial terutama *twitter* sangat besar sebagai sumber data yang menghasilkan informasi yang bermanfaat. Melalui *twitter* masyarakat memberikan opini berupa *tweet* terkait dengan JakLingko. Berbagai *tweet* yang diberikan dapat mengandung tanggapan yang positif, negatif, atau netral. Analisis terhadap opini inilah yang selanjutnya disebut dengan analisis sentimen.

Analisis sentimen merupakan data mining yang digunakan untuk menganalisis, memahami mengolah, dan mengekstrak data tekstual yang berupa opini terhadap entitas seperti produk, pelayanan, organisasi, individu, atau topik tertentu. Analisis sentimen merujuk pada penilaian yang

bersifat negatif atau positif (Pang & Lee, 2008). Opini tersebut dipisahkan ke kelas positif, negatif, atau netral menggunakan metode kalsifikasi *Support Vector Machine* (SVM). Dalam klasifikasi menggunakan SVM, peneliti diharuskan untuk menemukan fungsi pemisah antar kelas. *Margin* adalah jarak antara *hyperplane* dengan data terdekat pada masing-masing kelas. Bidang pembatas pertama membatasi kelas pertama dan bidang pembatas kedua membatasi kelas kedua sedangkan data yang berada pada bidang pembatas merupakan vektor-vektor yang terdekat dengan *hyperplane* terbaik disebut dengan *Support Vector* (Tan, Steinbach, & Kumar, 2006). Penelitian sebelumnya telah banyak dilakukan tentang analisis sentimen menggunakan data *twitter*. Penelitian yang dilakukan oleh Chory et. al (2018) tentang kepuasan publik terhadap layanan data operator telekomunikasi di Indonesia menggunakan SVM untuk mengklasifikasikan jenis sentimennya serta memanfaatkan pembobotan TF-IDF dalam meningkatkan akurasi. Rahat, A., Kahir, A. & Masum, A., (2019) melakukan analisis sentimen dengan membandingkan *Naïve Bayes* dan SVM sebagai metode klasifikasinya dan dihasilkan bahwa metode SVM mampu mengklasifikasikan sentimen lebih baik dibandingkan dengan *Naïve Bayes*. Penelitian tentang analisis sentimen terhadap program JakLingko juga telah dilakukan oleh Rachman, F. F., Nooraeni, R. & Yuliana, L., (2020). Penelitian tersebut menggunakan *tweet* dengan kata kunci masing-masing jenis transportasi yang terintegrasi dengan JakLingko yaitu MRT, LRT, KRL, dan TJ. Analisis sentimen dapat menghasilkan suatu evaluasi rangkuman berdasarkan opini publik tentang layanan transportasi umum di Jakarta. Data tersebut dapat mempermudah para *stakeholder* dalam memperbaiki layanan transportasi umum di Jakarta. Harapannya dengan dilakukannya analisis sentimen ini dapat diketahui permasalahan terhadap JakLingko, sehingga dapat meningkatkan minat warga DKI Jakarta untuk menggunakan transportasi umum.

METODOLOGI

Bahan dan Data

Data yang digunakan dalam penelitian ini adalah data *tweet* dari media sosial *twitter* dengan kata kunci "JakLingko". Proses pengambilan data dilakukan dalam periode 1 Juni 2022 - 31 Agustus 2022 menggunakan *scraper* di Python yaitu *snscrape*. Dari proses *scraping* yang dilakukan didapatkan data sebanyak 4308 *tweet* yang ditunjukkan pada TABEL 1.

TABEL 1. Hasil *Scraping Twitter*

Twitt ke-	*Datetime*	*Text*	*User*
1	31/08/2022	Asli. Apalagi mikrotrans JakLingko.... beh lama bangetðŸ˜©ðŸ˜	farhanadji
2	31/08/2022	@dickyrevolution @irafas_s @CNNIndonesia PANTESAN DIA GA TAU JAKLINGKO, ORANG PLANET BEKESONG RUPANYA	SudutMiliter
…	...	...	...
4307	01/06/2022	@injuniyaa nih pasti pas di kereta nnti padett bgt, eh tapi kl dari tanabang ke kota bisa naik angkot 08 atau JakLingko no 10 tau sebenernya	kabogohyuta
4308	01/06/2022	@PT_Transjakarta Pagi, saya pengguna JakLingko dengan adanya JakLingko saya sanggat terbantu,cuman sanggat di sayangkan supir jak 17(pologadung-senen) tolong di tindak lanjutin saya sudah 3hari berturut " memberhentikan mobil di busstop dengan mobil yang sama dan pengemudi yang sama tdk berhenti	rizkyasyam94

Metode Penelitian

Text Preprocessing

Data *twitter* adalah bentuk dari *text mining* yaitu kumpulan data tekstual yang tidak terstruktur. Dari data yang tidak terstruktur ini akan diidentifikasi apakah ada pola yang menarik dari dokumen tersebut (Feldman & Sanger, 2007). Dokumen pada umumnya memiliki bentuk yang tidak terstruktur. Oleh karena itu perlu dilakukan suatu proses yang dapat mengubah bentuk data yang sebelumnya tidak terstruktur ke dalam bentuk data yang terstruktur. Tahapan proses tersebut disebut dengan *preproccesing* yang meliputi (Faret & Reitan, 2015):

1. *Casefolding*

Casefolding merupakan proses dalam *text preprocessing* yang dilakukan untuk menyeragamkan karakter pada dokumen. Pada proses *casefolding* dilakukan pengubahan huruf kapital menjadi huruf kecil dalam dokumen. Pengubahan tersebut dilakukan dari huruf "a" sampai dengan "z".

2. *Tokenizing*

Tokenizing secara garis besar dapat didefinisikan sebagai peoses pemecahan sekumpulan karakter dalam suatu teks ke dalam satuan kata. *Tokenizing* merupakan proses pemisahan suatu rangkaian karakter berdasarkan karakter spasi, dan mungkin pada waktu yang bersamaan dilakukan juga proses penghapusan karakter tertentu, seperti tanda baca.

3. *Filtering*

Filtering merupakan tahapan pembuangan kata-kata yang dianggap tidak penting dan tidak berpengaruh terhadap makna kata. Tahap filtering merupakan tahapan penyeleksian kata-kata penting dari hasil tokenizing. Tahap *filtering* ini menggunakan daftar *stopword*.

4. *Normalization*

Normalization merupakan proses mengubah format teks untuk tujuan tertentu. Pendekatan normalisasi dibagi menjadi 2 kelompok. Yang pertama menggunakan informasi kontekstual untuk menerjemahkan bahasa non-standar ke dalam bahasa standar, dan yang kedua mengganti kata-kata berbasis leksikal dengan bentuk yang sesuai dalam bahasa standar.

5. *Stemming*

Stemming didefinisikan sebagai proses membentuk suatu kata menjadi kata dasarnya dengan penghilangan awalan dan akhiran kata.

Term Frequency – Inverse Document Frequency (TF-IDF)

TF-IDF adalah suatu cara untuk memberikan bobot hubungan suatu kata atau term terhadap suatu dokumen. Algoritma ini menggabungkan dua konsep untuk perhitungan bobot, yaitu frekuensi kemunculan sebuah kata di dalam sebuah dokumen tertentu atau TF dan invers dari frekuensi dokumen yang mengandung kata tersebut atau IDF. Perhitungan TF-IDF sendiri dilakukan dengan menggunakan persamaan (Salton & Buckley, 1988):

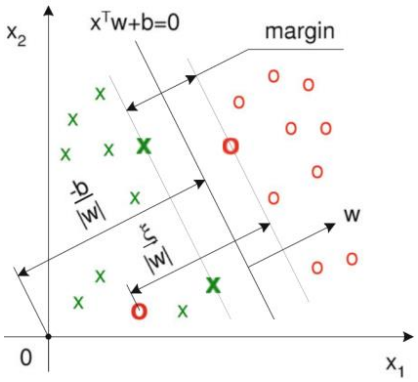
$$W_{j,i} = \frac{n_{j,i}}{\sum_k n_{k,i}} \log_2 \frac{D}{d_j} \quad (1)$$

dengan,

$W_{j,i}$: pembobotan TF-IDF untuk term ke *j* pada dokumen ke *i*
$n_{j,i}$: jumlah kemunculan term ke *j* pada dokumen ke *i*
$\sum_k n_{k,i}$: jumlah kemunculan seluruh *term* pada dokumen ke *i*
D : banyaknya dokumen
d_j : banyaknya dokumen yang mengandung *term* ke *j*

Support Vector Machine

Support Vector Machine (SVM) merupakan suatu teknik untuk menemukan *hyperplane* yang bisa memisahkan dua set data dari dua kelas yang berbeda (Cortes & Vapnik, 1995). *Hyperplane* terbaik adalah *hyperplane* yang memiliki *margin* yang paling lebar. *Margin* (m) adalah jarak antara *hyperplane* dengan data terdekat pada masing-masing kelas. *Margin* pertama membatasi kelas pertama dan *margin* kedua membatasi kelas kedua sedangkan data yang berada pada bidang pembatas merupakan vektor-vektor yang terdekat dengan *hyperplane* terbaik disebut dengan *Support Vector* (Tan, Steinbach, & Kumar, 2006).



GAMBAR 1. Hyperplane Pemisah dan Margin (Diperoleh dari Hardle & Simar, 2015)

GAMBAR 1 menunjukkan pemisahan dua kelompok data oleh suatu fungsi *hyperplane*. Data pada kelompok pertama ditunjukkan dengan simbol (x) warna hijau, sedangkan data dari kelompok kedua ditunjukkan dengan simbol (o) berwarna merah. Terlihat bahwa fungsi *hyperplane* memisahkan kedua kelompok data. Garis sejajar yang ada di kanan dan kiri dari *hyperplane* merupakan *margin*. Data yang jatuh pada margin disebut sebagai *support vector* yang menentukan fungsi *hyperplane* yang akan terbentuk.

Klasifikasi linear *hyperplane* SVM dilambangkan dengan:

$$f(x) = w^T x + b \tag{2}$$

Sehingga akan diperoleh persamaan:

$$[w^T . x_i + b] \geq 1 \; untuk \; y_i = +1 \tag{3}$$

$$[w^T . x_i + b] \leq -1 \; untuk \; y_i = -1 \tag{4}$$

Untuk x_i =himpunan data *training*, $i = 1,2,\ldots,n$, dan y_i = label kelas dari x_i

Untuk mendapatkan *hyperplane* terbaik, perlu ditemukan *hyperplane* di tengah antara dua batas kelas. Mendapatkan *hyperplane* terbaik setara dengan memaksimalkan *margin* atau jarak antara dua kelompok objek dari kelas yang berbeda. *Hyperplane* terbaik didapatkan dengan metode *Quadratic Progamming* (QP) yaitu dengan meminimalkan $w^T w$ dengan fungsi batasan $y_i(w^T . x_i + b) \geq 1$, $i = 1,2,3,\ldots,n$. Persoalan optimasi tersebut diselesaikan dengan membentuk formula *lagrange* sebagai berikut.

$$L_p(\mathbf{w}, \mathrm{b}, \alpha) = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{n} \alpha_i \left[y_i(\mathbf{x}_i'\mathbf{w} + b) - 1 \right] \tag{5}$$

dengan α adalah *lagrange multiplier* pada *lagrangian primal problem* (5). Nilai optimal dari persamaan (5) didapatkan dengan meminimumkan $L_p(\mathbf{w},b,\alpha)$ terhadap **w** dan b sehingga didapatkan *lagrangian* untuk dual problem sebagai berikut.

$$L_D(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \mathbf{x}_i' \mathbf{x}_j \quad (6)$$

Solusi dari masalah pemrograman kuadratik yaitu nilai α didapatkan dengan memaksimalkan fungsi L_D (α) dengan batasan $\alpha_i \geq 0, \quad \sum_{i=1}^{n} \alpha_i y_i = 0$. Kemudian ditentukan kelas dari data *training* dengan fungsi berikut.

$$g(x) = sign(\mathbf{x}_i' \mathbf{w} + b) \quad (7)$$

dengan $\mathbf{w} = \sum_{i=1}^{n} a_i y_i x_i$ dan $b = \frac{1}{2}(x_{+1} + x_{-1})w$. x_{+1} dan x_{-1} adalah dua *Support Vector* yang mengikuti kelas yang berbeda untuk $y(\mathbf{x}'\mathbf{w} + b) = 1$. Data *training* dengan i > 0 terletak pada *hyperplane* disebut dengan *support vector* dan memiliki nilai $\alpha_i \geq 0$ sedangkan untuk yang lain $\alpha_i = 0$ atau tidak memberikan nilai terhadap fungsi *hyperplane*.

SVM pada dasarnya menggunakan fungsi linier sebagai pemisah antara dua kelompok data. Namun SVM mampu dikembangkan menjadi sebuah adalah algoritma yang mampu mengubah data *training* ke dimensi yang lebih tinggi menggunakan pemetaan nonlinier. Konsep ini disebut dengan *kernel trick*. Fungsi *kernel* akan memetakan data ke dimensi yang lebih tinggi. Beberapa fungsi pembentuk matriks kernel yang umum digunakan pada SVM adalah:

1. Kernel *Linear*
   $$K(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i' \mathbf{x}_j \quad (8)$$
2. Kernel Polinomial
   $$K(\mathbf{x}_i, \mathbf{x}_j) = (\delta \mathbf{x}_i' \mathbf{x}_j + r)^p, \delta > 0 \quad (9)$$
3. Kernel *Radial Basis Function* (RBF)
   $$K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\gamma \left\| \mathbf{x}_i - \mathbf{x}_j \right\|^2\right), \gamma > 0 \quad (10)$$
4. Kernel Sigmoid
   $$K(\mathbf{x}_i, \mathbf{x}_j) = \tanh(\delta \mathbf{x}_i' \mathbf{x}_j + r) \quad (11)$$

Pemilihan fungsi kernel yang tepat merupakan hal yang sangat penting karena akan menentukan *feature* space dimana fungsi *classifier* akan dicari. Sepanjang fungsi kernelnya sesuai, SVM akan beroperasi secara benar meskipun pemetaan yang digunakan tidak diketahui (Santosa, 2007).

K-Fold Cross Validation

Cross validation adalah metode statistik untuk mengevaluasi dan membandingkan algoritma pembelajaran dengan membagi data menjadi dua bagian yaitu data *training* yang digunakan untuk *training* dan data *testing* yang digunakan untuk memvalidasi model (Refaeilzadeh, Tang, dan Liu, 2008). *K-fold cross validation* akan membagi data ke dalam k subset yang saling bebas yaitu S_1, S_2, ..., S_k dengan jumlah pengamatan tiap subset hampir sama, selanjutnya jika satu subset menjadi data *testing* maka *k-1* subset yang akan menjadi data *training* (Han, Kamber, & Jian, 2006). Metode ini digunakan dengan tujuan menghilangkan bias yang ada pada data, dimana proses pelatihan dan pengujian data dilakukan sebanyak *k*-kali. Kemudian *k*-1 bagian akan digunakan sebagai data *training* dan satu bagian sisanya sebagai data *testing* untuk validasi model. Nilai *Mean Square Error* (*MSE*) akan dihitung untuk melihat *error* pada model. Nilai tersebut dapat dihitung dengan rumus:

$$MSE = \sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n} \quad (12)$$

Setelah proses perulangan sebanyak *k*-kali, nilai *mean square of error* setiap perulangan akan dihitung sehingga akan diperoleh nilai $MSE1$, $MSE2$,… , $MSEk$. Model terbaik akan ditentukan berdasarkan *performance metric model* yakni menggunakan rata-rata MSE dari setiap iterasi.

HASIL DAN PEMBAHASAN

Text Preprocessing

Semua teks *tweet* selanjutnya memasuki tahap *preprocessing* yang mencakup *casefolding, tokenizing, filtering, normalization,* dan *stemming.* Keseluruhan proses ini diperlukan agar data teks nantinya lebih mudah untuk diolah dan mampu menghasilkan klasifikasi sentimen yang baik dan tepat.

Proses yang pertama kali dilakukan ialah *casefolding*, di mana seluruh huruf yang berada seluruh *tweet* cuitan akan diubah kedalam bentuk huruf kecil. Misalkan saja proses *casefolding* dilakukan pada cuitan ke-94, yakni "@dsimbol1 @aniesbaswedan @c40cities @DKIJakarta @uclg_org @mrtjakarta Yg sulit itu bukan menciptakan, tp menjaga, merawat dan mengintegrasikan. Skrg turun stasiun ada JakLingko, turun halte ada JakLingko, udah jaminan sampe deket rumah terintegrasi ", maka semua huruf kapital yang terdapat dalam kalimat tersebut akan diubah menjadi huruf kecil. Hasil dari proses tersebut adalah "@dsimbol1 @aniesbaswedan @c40cities @dkijakarta @uclg_org @mrtjakarta yg sulit itu bukan menciptakan, tp menjaga, merawat dan mengintegrasikan. skrg turun stasiun ada jaklingko, turun halte ada jaklingko, udah jaminan sampe deket rumah terintegrasi".

Tahap selanjutnya yakni *tokenizing* dimana setiap kalimat hasil proses *casefolding* dibersihkan dari tanda baca, angka, spasi berlebihan, hingga *emoticon*. Kalimat-kalimat ini kemudian dipecah menjadi beberapa kata. Jika menggunakan cuitan ke-94 sebagai contoh hasil proses *tokenizing*, akan diperoleh 26 kata sebagai hasil dari proses ini, yakni 'org', 'yg', 'sulit', 'itu', 'bukan', 'menciptakan', 'tp', 'menjaga', 'merawat', 'dan', 'mengintegrasikan', 'skrg', 'turun', 'stasiun', 'ada', jaklingko, 'turun', 'halte', 'ada', jaklingko, 'udah', 'jaminan', 'sampe', 'deket', 'rumah', dan 'terintegrasi'.

Teks berikutnya memasuki proses *filtering*, di mana setiap *stopword* dihilangkan dari susunan kalimat, sehingga diharapkan setiap katanya merupakan kata yang memiliki makna. Sebagai contoh jika proses *filtering* dilakukan pada 16 kata hasil proses *tokenizing* cuitan ke-94, hanya akan tersisa 20 kata, yakni 'org', 'sulit', 'menciptakan', 'tp', 'menjaga', 'merawat', 'mengintegrasikan', 'skrg', 'turun', 'stasiun', jaklingko, 'turun', 'halte', jaklingko, 'udah', 'jaminan', 'sampe', 'deket', 'rumah', 'terintegrasi'.

Proses *normalization* dilakukan setelah melakukan *filtering*. Pada tahap ini, setiap kata akan diubah ke kata lain yang sifatnya sama. Hal seperti ini biasanya dilakukan karena adanya bahasa tidak formal atau singkatan yang digunakan dalam sebuah cuitan. Sebagai contoh, kata 'org', 'tp', 'skrg', 'udah', 'sampe', 'deket' pada cuitan ke-94 diubah menjadi kata 'orang', 'tapi', 'sekarang', 'sudah', 'sampai', 'dekat'

Proses terakhir dalam tahap *text preprocessing* adalah *stemming*. Melalui proses ini, semua kata hasil proses *normalization* akan diubah ke dalam bentuk kata dasarnya. Sebagai contoh, kata 'menciptakan', 'menjaga', 'mengintegrasikan', dan 'terintegrasi' pada *tweet* ke-94 diubah menjadi 'cipta', 'jaga', 'rawat', dan 'integrasi' . Keseluruhan proses *text preprocessing* pada *tweet* ke-94 dapat dilihat pada TABEL 2.

TABEL 2. Hasil *Text Preprocessing* data *tweet* ke-94

Teks Asli	Proses	Hasil		
@dsimbol1 @aniesbaswedan @c40cities @DKIJakarta @uclg_org @mrtjakarta Yg sulit itu bukan menciptakan, tp menjaga, merawat dan mengintegrasikan. Skrg turun stasiun ada jaklingko,	*Casefolding*	@dsimbol1 @aniesbaswedan @c40cities @dkijakarta @uclg_org @mrtjakarta yg sulit itu bukan menciptakan, tp menjaga, merawat dan mengintegrasikan. skrg turun stasiun ada jaklingko, turun halte ada jaklingko, udah jaminan sampe deket rumah terintegrasi		
	Tokenizing	org yg sulit itu	dan mengintegrasikan skrg turun	ada JakLingko udah jaminan

turun halte ada jaklingko, udah jaminan sampe deket rumah terintegrasi		bukan menciptakan tp menjaga merawat	stasiun ada jaklingko turun halte	sampe deket rumah terintegrasi
	Filtering	org sulit menciptakan tp menjaga merawat mengintegrasikan	skrg turun stasiun jaklingko turun halte JakLingko	udah jaminan sampe deket rumah terintegrasi

Teks Asli	**Proses**	**Hasil**		
@aniesbaswedan @c40cities @DKIJakarta @uclg_org @mrtjakarta Yg sulit itu bukan menciptakan, tp menjaga, merawat dan mengintegrasikan. Skrg turun stasiun ada jaklingko, turun halte ada jaklingko, udah jaminan sampe deket rumah terintegrasi	*Normalization*	orang sulit menciptakan tapi menjaga merawat mengintegrasikan	sekarang turun stasiun jaklingko turun halte jaklingko	sudah jaminan sampai dekat rumah terintegrasi
	Stemming	orang sulit cipta tapi jaga rawat integrasi	sekarang turun stasiun jaklingko turun halte jaklingko	sudah jaminan sampai dekat rumah integrasi

Pembobotan *Term frequency-inverse document frequency* (TF-IDF)

Dokumen mewakili masing-masing data *tweet* yang diteliti sehingga dalam penelitian ini terdapat 4308 dokumen. Dalam *Natural Languange Process* (NLP) dokumen tersebut tidak bisa langsung dianalisis, dibutuhkan *preprocessing* yang telah dilakukan pada tahap sebelumnya, dan pembobotan menggunakan *Term frequency-inverse document frequency* (TF-IDF).

Tabel 3. Hasil Pembobotan TF-IDF

Indeks Dokumen	**Indeks Kata**	**Nilai TF-IDF**
0	2443	0,037
0	217	0,486
…	…	…
0	2673	0,221
1	5059	0,444
1	547	0,519
…	…	…
1	2443	0,06
…	…	…

4307	854	0,341
…	…	…
4307	2443	0,078

TABEL 3 menunjukkan hasil pembobotan TF-IDF. Pada tabel tersebut terlihat bahwa indeks dokumen dimulai dari 0 sampai 4307 yang menunjukkan urutan dari *tweet* yang diteliti, Indeks kata merupakan kode untuk setiap kata yang ada pada keseluruhan data *tweet*. Pada penelitian ini terdapat total 6498 kata sehingga indeks kata akan dimulai dari 0 sampai 6497. *Tweet* pertama atau dokumen ke-0 mengandung kata dengan kode indeks 2443 sampai 2673. Nilai TF-IDF menunjukkan bobot dari setiap kata yang muncul pada setiap dokumen. Kata yang muncul berkali-kali dalam sebuah dokumen, namun tidak muncul berkali-kali di dokumen lain, akan memiliki bobot yang tinggi. Sedangkan kata yang muncul berkali-kali tidak hanya pada dokumen tersebut tapi juga sering muncul pada dokumen yang lain maka kata tersebut dianggap tidak terlalu berarti bagi dokumen itu sehingga bobot TF-IDF yang didapatkan rendah. Apabila ditinjau dari akumulasi nilai bobot TF-IDF untuk masing-masing kata maka akan didapatkan kata dengan total TF-IDF yang tertinggi sampai yang terendah. Kata dengan jumlah TF-IDF tinggi artinya memiliki bobot yang tinggi pada setiap dokumen yang mengandung kata tersebut. Sedangkan kata dengan total TF-IDF rendah berarti memiliki bobot yang rendah pada setiap dokumen yang mengandung kata tersebut. Hasil total nilai bobot TF-IDF ditunjukkan pada TABEL 4.

TABEL 4. Hasil Pembobotan TF-IDF

Kata ke-	Kata	Jumlah TF-IDF
2458	JakLingko	320.256084
6064	transjakarta	164.304477
2749	kartu	157.291895
213	angkot	107.137189
…	…	…
1061	collaboration	0.204733
5140	samawa	0.204733
5572	sport	0.204733
4358	panah	0.204733

Berdasarkan TABEL 4, kata JakLingko, transjakarta, dan kartu mempunyai nilai yang tinggi. Artinya ketiga kata tersebut mempunyai bobot TF-IDF yang tinggi pada setiap dokumen yang mengandung kata tersebut. Selanjutnya, hasil perhitungan TF-IDF akan dibentuk kedalam matriks pembobotan kata berukuran 6498 x 4308 yang berisi nilai TF-IDF setiap kata untuk setiap dokumen. Matriks TF-IDF inilah yang kemudian digunakan untuk analisis lebih lanjut yaitu klasifikasi sentimen terhadap JakLingko menggunakan metode SVM.

Support Vector Machine (SVM)

Setelah didapatkan matriks TF-IDF kemudian dilanjutkan dengan klasifikasi. Sebelum dilakukan klasifikasi, terlebih dahulu dilakukan *labelling* untuk tiap dokumen atau *tweet*. Label yang diberikan yaitu positif untuk *tweet* yang mengandung sentimen positif terhadap JakLingko, negatif untuk *tweet* yang berisi tanggapan negatif tentang JakLingko dan netral untuk *tweet* yang tidak mengandung sentimen positif atau negatif, biasanya berupa info atau pertanyaan dari pengguna *twitter*. Parameter yang digunakan adalah parameter *cost* (*C*) dengan nilai 2^0, 2^1 , 2^2 , 2^3, 2^4 dan parameter RBF kernel yaitu *gamma* (*γ)* dengan nilai 2^0 , 2^{-1}, 2^{-2} , 2^{-3}, 2^{-4}. Pengujian performa klasifikasi SVM dilakukan dengan metode 5-*fold cross-validation* pada masing-masing kombinasi nilai parameter C dan γ sehingga didapatkan nilai parameter C dan γ yang paling optimal masing-masing adalah 2^0 dan 2^{-4} . Evaluasi model pada setiap *fold* dalam proses c*ross validation* untuk parameter C = 2^0 dan $\gamma = 2^{-4}$ didapatkan nilai akurasi masing-masing yaitu 0,72753623, 0,72119013, 0,69150943, 0,71683599, dan

0,70264151. Rata-rata akurasi yang didapatkan dari kelima *fold* tersebut adalah 0,711. Berdasarkan hasil tersebut dapat disimpulkan bahwa rata-rata akurasi dari kelima *fold* sudah baik, selain itu nilainya juga stabil.

Fungsi *hyperplane* yang terbentuk dari model terbaik SVM dengan parameter $C = 2^0$ dan $\gamma = 2^{-4}$ adalah $g(x) = sign(\mathbf{x}_i'\mathbf{w} + b)$ dengan $w = \sum_{i=1}^{345}\sum_{j=1}^{345}\alpha_i y_i K(x_i, x_j)$ dan $K = exp\left(-2^{-4}\|x_i - x_j\|^2\right)$ dan b adalah suatu konstan dimana b = -0.048. Model tersebut akan diterapkan pada data *testing* untuk diketahui bagaimana performa klasifikasinya. TABEL 5 menunjukkan hasil prediksi menggunakan model terbaik yang dibandingkan dengan nilai sentimen sesungguhnya.

TABEL 5. Hasil Prediksi

Tweet	**Sentimen**	**Prediksi**
@askvicong @hericz @Outstandjing Aku baru naik JakLingko, dan ternyata sampai malam! Harusnya ada penitipan sepeda di pool/pemberhentian Jak Lingko. Kemarin, ikut nitip aja di pool :))	Positif	Negatif
@CaloTerminl Ohh dia lagi toh Yg dulu pernah gua kritik kalo angkot JakLingko ada yg ugal2an malah gak terima😂😂😂	Negatif	Positif
@alwaysbeliozz @catteyes11 @convomf cara bikin kartu JakLingko gmn kak?	Netral	Netral
@Sanwo18 Pak Rachmat, naik JakLingko pulang dari Nice-so, malam ini 😍 🎗JakLingko Cibubur, JAKTIM (BUKAN Depok, Bekasi apalagi Bogor)	Positif	Positif
@whiskersperson Hai kak. Untuk menggunakan layanan Mikrotrans silakan menggunakan kartu berlogo JakLingko. Terima kasih ^SJ	Netral	Netral
@PT_Transjakarta Kalau pake kartu JakLingko bisa?? Tlg dijawab yaa..sy udh nanya 2x	Negatif	Netral
@passaintt @zahratalitha_ sorry nimbrung, setau gue sih cuma butuh tiket gelang fisik dari jakmania per-korwil. jadi harus cocok tiket gelang utara fisik sama JakLingko utara jakmania, gitu contohnya.	Netral	Netral
…	…	…
Antrian JakLingko tujuan Tenabang - Kota.. Tambah Armada Please Min @JakLingkojkt cc : @DKIJakarta @aniesbaswedan https://t.co/km5iZBFnZU	Negatif	Negatif
@worksfess Atau sender bisa beli kartu JakLingko yang vending machinenya tersedia di beberapa halte, kartu JakLingko bisa buat mrt, lrt, krl juga	Netral	Netral

Selanjutnya hasil klasifikasi dari seluruh *tweet* pada data *training* dirangkum dalam *confusion matrix* yang ditunjukkan pada TABEL 6.

TABEL 6. *Confusion Matrix* Data *Testing*

Kelas Aktual	**Kelas Prediksi**			**Total**
	Negatif	**Netral**	**Positif**	
Negatif	154	43	37	234
Netral	44	254	49	347
Positif	36	47	198	281
Total	234	344	284	862

Berdasarkan rangkuman klasifikasi yang diberikan pada TABEL 6, terlihat bahwa model mampu mengklasifikasikan *tweet* dengan sentimen negatif dengan benar sebanyak 154 *tweet*, sentimen positif sebanyak 198 *tweet* dan *tweet* netral dapat diklasifikasikan dengan benar sebanya 254 *tweet*.

Pengukuran seberapa baik model bekerja dengan melihat jumlah *tweet* yang berhasil dengan benar diprediksi dapat dinyatakan dalam nilai akurasi sebagai berikut.

$$\text{akurasi} = \frac{\sum True\ Positive}{n_{testing}} = \frac{154 + 254 + 198}{862} = \frac{606}{862} = 0{,}703$$

Penerapan model SVM terbaik pada data *training* mampu mencapai nilai akurasi 0,703.

Gambaran Sentimen Positif dan Negatif

Wordcloud dapat memberikan gambaran tentang kata-kata yang sering disebutkan dalam *tweet*. Topik utama yang menjadi pembahasan dari *tweet* yang mengandung sentimen positif dan sentimen negatif ditunjukkan pada GAMBAR 2. Kata yang lebih sering disebutkan akan berukuran lebih besar dari yang lain, semakin jarang kata tersebut muncul, maka ukurannya akan semakin kecil.

(a) **(b)**

GAMBAR 2. *Wordcloud* Sentimen: (a) Positif, (b) Negatif

Berdasarkan *wordcloud* pada GAMBAR 2, JakLingko dan Jakarta sama-sama menjadi topik yang sering disebutkan pada *tweet* yang mengandung sentimen baik positif maupun negatif. Namun pada *tweet* positif GAMBAR 2 (a), muncul kata-kata yang lain seperti angkot dan gratis. Sedangkan pada tweet negatif GAMBAR 2 (b) terdapat kata yang juga banyak disebut yaitu transjakarta, angkot dan kartu JakLingko. Kata angkot dan gratis muncul karena banyak *tweet* positif yang menyatakan bahwa adanya tarif angkot gratis sangat memudahkan masyarakat. Sedangkan transjakarta sering disebut oleh pengguna *twitter* saat menyampaikan keluhan tentang JakLingko, hal ini dikarenakan akun *twitter* JakLingko yang kurang aktif sehingga masyarakat lebih memilih untuk menyebut akun transjakarta saat menyampaikan keluhan. Angkot dan kartu juga banyak disebut pada *tweet* negatif karena terdapat pelanggan yang terkendala dengan penggunaan kartu JakLingko dan penggunaan angkot yang walaupun sudah terintegrasi dengan JakLingko namun pelayanannya masih kurang baik.

KESIMPULAN DAN SARAN

Klasifikasi data *tweet* JakLingko ke dalam sentimen positif dan negatif didapatkan rata-rata akurasi pada data *training* melalui proses *cross validation* sebesar 0,711. Nilai tersebut termasuk baik dan akurasinya stabil pada tiap *fold*. Akurasi yang diterapkan pada data *training* didapatkan nilai yang tidak jauh berbeda yaitu 0,703. Topik yang banyak dibicarakan pada sentimen positif yaitu tentang tarif angkot jaklingko gratis yang sangat memudahkan masyarakat. Pada sentimen negatif, pengguna

twitter menyampaikan keluhan dengan ikut melakukan *mention* terhadap akun @PT_Transjakarta dan mengeluhkan tentang penggunaan kartu JakLingko serta angkot yang masih belum baik pelayanannya.

Saat ini terdapat banyak pengembangan metode *text preprocessing*. Penerapan *text preprocessing* yang lebih baik, detail, dan teliti diharapkan mampu meningkatkan nilai akurasi dalam klasifikasi data sentimen.

REFERENSI

Chory, R. N., Nasrun, M. & Setianingsih, C., 2018. Sentiment Analysis on User Satisfaction Level of Mobile Data Services Using Support vector machine (SVM) Algorithm. *The 2018 IEEE IoTaIS.*

Cortes, C. & Vapnik, V., 1995. Support-Vector Networks. *Machine Leaming,* Volume 20, pp. 273-297.

DataReportal, 2022. *Digital in Indonesia 2022.* [Online] Available at: https://datareportal.com/reports/digital-2022-indonesia [Accessed 1 Juli 2022].

Faret, J. & Reitan, J., 2015. *Twitter Sentiment Analysis: Exploring the Effects of Linguistic Negation.* Norwegia: Norwegian University of Science and Technology.

Feldman, R. & Sanger, J., 2007. *The Text Mining Handbook.* New York: Cambridge University Press.

Han, J., Kamber, M. & Jian, P., 2006. *Data Mining: Concept and Techniques.* 3th ed ed. San Fransisco: Morgan Kaufmaan.

Hardle, W. & Simar, L., 2015. *Applied Multivariate Statistical Analysis.* Verlag Berlin Heidelberg: Springer.

Pang, B. & Lee, L., 2008. Opinion Mining and Sentiment Analysis. *Foundation and Trends in Information Retrieval, https://doi.org/10.3748/wjg.v22.i45.9898.*

Rachman, F. F., Nooraeni, R. & Yuliana, L., 2020. *Public opinion of transportation integrated (JakLingko) in DKI Jakarta Indonesia.* s.l., 5th International Conference on Computer Science and Computational Intelligence 2020.

Rahat, A., Kahir, A. & Masum, A., 2019. *Comparison of Naive Bayes and SVM Algorithm based on Sentiment Analysis Using Review Dataset.* s.l., Proc. 2019 8th Int. Conf. Syst. Model. Adv. Res. Trends, SMART 2019, pp. 266–270, 2020, doi: 10.1109/SMART46866.2019.9117512.

Refaeilzadeh, P., Tang, L. & Liu, H., 2008. *Cross Validation.* s.l.:Arizona State University.

Salton, G. & Buckley, C., 1988. Term-Weighting Approaches in Automatic Text Retrieval. *Jurnal Information Processing and Management,* Volume Vol.24, No. 5, pp. 512-523.

Santosa, B., 2007. *Data Mining: teknik Pemanfaatan Data untuk Keperluan Bisnis, teori da Aplikasi.* s.l.:Graha Ilmu.

Simanjuntak, E., 2021. *The effect of brand personality dimensions on self congruity and functional congruity on brand attitude: A study of Jaklingko users.* Jakarta, Proceedings of the International Seminar of Contemporary Research on Business and Management (ISCRBM 2021).

Tan, P., Steinbach, M. & Kumar, V., 2006. *Introduction to Data Mining (4th ed.).* Boston: Pearson Addison Wesley.

Triana, S., Sjafruddin, A., Karsaman, R. H. & Kaderi, S., 2022. *Integration Of Mass Public Transport Fare In The Jakarta Area.* s.l., IOP Conference Series: Earth and Environmental Science.

Unit Pelayanan Statistik, 2019. *Survei Evaluasi Layanan Transportasi Terintegrasi.* [Online] Available at: https://statistik.jakarta.go.id/media/2020/06/Buku-Survei-Evaluasi-Layanan-Transportasi-Terintegrasi-Jak-Lingko-2019.pdf [Accessed 1 Mei 2022].