

ANALISIS SENTIMEN APLIKASI GOJEK MENGGUNAKAN SUPPORT VECTOR MACHINE DAN K NEAREST NEIGHBOR

M. Nurul Muttaqin[✉], Iqbal Kharisudin

Jurusan Matematika, FMIPA, Universitas Negeri Semarang, Indonesia
Gedung D7 Lt. 1, Kampus Sekaran Gunungpati, Semarang, 50229

Info Artikel

Sejarah Artikel:
Diterima Juli 2021
Disetujui November 2021
Dipublikasikan November 2021

Keywords:
*Sentiment Analysis,
Support Vector Machine,
K Nearest Neighbor*

Abstrak

Gojek merupakan perusahaan decacorn dari Indonesia yang berkembang sangat pesat. Dengan semakin banyaknya pengguna aplikasi Gojek membuat perusahaan harus senantiasa menjaga dan meningkatkan mutu pelayanannya. Analisis sentimen merupakan solusi untuk mengekstrak sebuah sumber informasi dalam bentuk teks dengan jumlah data yang besar dalam waktu yang singkat. Tujuan penelitian ini adalah melakukan klasifikasi pada ulasan Aplikasi Gojek menggunakan metode *Support Vector Machine (SVM)* dan *K Nearest Neighbor (KNN)*. Sumber data terbaik untuk mendapatkan dataset dalam penelitian analisis sentimen adalah Google Playstore, dikarenakan data yang diperoleh lebih bersih dan tidak mengandung unsur iklan ataupun promosi. Diperoleh hasil akurasi menggunakan Metode KNN dengan nilai $K=22$ memperoleh nilai akurasi, presisi, dan recall berturut-turut sebesar 82,14%, 82,28%, dan 95,43%, sedangkan metode SVM dengan kernel *linear* dan parameter $C=1$ memperoleh nilai akurasi, presisi, dan recall berturut-turut sebesar 87,98%, 88,55%, dan 95,43%. Dapat disimpulkan bahwa metode SVM melakukan klasifikasi dengan lebih baik dibandingkan metode KNN dalam melakukan analisis sentimen pada ulasan aplikasi Gojek di Google Playstore.

Abstract

Gojek is a decacorn company from Indonesia that is growing very rapidly. With the increasing number of Gojek application users, companies must always maintain and improve the quality of their services. Sentiment analysis is a solution for extracting a source of information in the form of text with a large amount of data in a short time. The purpose of this study is to classify the Gojek Application review using the Support Vector Machine (SVM) and K Nearest Neighbor (KNN) methods. The best data source for obtaining datasets in sentiment analysis research is Google Playstore, because the data obtained is cleaner and does not contain advertising or promotional elements. Accuracy results obtained using the KNN method with a value of $K=22$ obtained accuracy, precision, and recall values of 82.14%, 82.28%, and 95.43%, respectively, while the SVM method with linear kernel and parameter $C=1$ obtained the values of accuracy, precision, and recall of 87.98%, 88.55%, and 95.43%, respectively. It can be concluded that the SVM method performs classification better than the KNN method in conducting sentiment analysis on Gojek application reviews on Google Playstore.

How to cite:

Muttaqin, M. N., Kharisudin, I. 2021. Analisis Sentimen Pada Ulasan Aplikasi Gojek Menggunakan Metode Support Vector Machine dan K Nearest Neighbor. *UNNES Journal of Mathematics*. 10(2):22-27.

© 2021 Universitas Negeri Semarang

[✉]Alamat korespondensi:

E-mail: mnurulmuttaqin@students.unnes.ac.id

p-ISSN 2252-6943

e-ISSN 2460-5859

PENDAHULUAN

Perkembangan teknologi di Indonesia memunculkan banyaknya aktivitas manusia yang dapat didukung dengan sarana digital, salah satunya dalam bidang transportasi. Hal ini terlihat dari bermunculannya berbagai model transportasi berbasis online pada kota-kota besar di Indonesia. Salah satu transportasi online yang paling diminati masyarakat di Indonesia adalah Gojek. PT Aplikasi Karya Anak Bangsa (Gojek) merupakan sebuah perusahaan decacorn asal Indonesia yang didirikan pada tahun 2010 di Jakarta oleh Nadiem Makarim. Dalam menjalankan bisnisnya, PT Aplikasi Karya Anak Bangsa (Gojek) membuat suatu aplikasi yang didaftarkan pada berbagai platform yang terintegrasi dengan smartphone, salah satunya yaitu Google Playstore.

Seiring meningkatnya jumlah pengguna smartphone dan kendaraan berakibat pada semakin banyaknya para pengguna aplikasi gojek, hal ini membuat tuntutan terhadap mutu pelayanan aplikasi juga meningkat. Dalam Google Playstore setiap pengguna dapat memberikan rating dan ulasan terhadap suatu aplikasi. Menurut (Wayan & Saraswati, 2013), ulasan dan informasi suatu produk disimpan dalam bentuk teks, maka *text mining* merupakan solusi dalam pengambilan informasi yang berbentuk teks. Salah satu proses pengambilan informasi *text mining* dalam kategori teks adalah *Sentiment Analysis* atau Analisis Sentimen. Analisis sentimen adalah studi komputasi dari opini-opini, sentimen, serta emosi yang diekspresikan dalam teks (Liu, 2016). Analisis sentimen diperlukan sebagai alat untuk mengklasifikasikan sebuah informasi yang berbentuk teks bahasa kedalam kategori positif dan negatif (Hidayat, 2015). Analisis sentimen digunakan untuk melihat kecenderungan suatu sentimen atau pendapat, apakah pendapat tersebut cenderung beropini positif atau negatif.

Penelitian sebelumnya yang dilakukan oleh (Kurniawan, 2017) dengan judul 'Implementasi Text Mining pada Analisis Sentimen Pengguna Twitter Terhadap Media Mainstream Menggunakan Naïve Bayes Classifier dan Support Vector Machine' diperoleh hasil bahwa metode *Support Vector Machine* melakukan klasifikasi dengan lebih baik dibandingkan Naïve Bayes Classifier dengan tingkat akurasi sebesar 97,9%. Sedangkan penelitian lain yang dilakukan oleh (Lestari, 2019) dengan judul 'Perbandingan

Klasifikasi Beasiswa Toyota Astra Menggunakan K-Nearest Neighbor Classifier dan Naive Bayes Sebagai Penentu Metode Klasifikasi Pada Sistem Pendukung Keputusan Penerimaan Beasiswa Toyota Astra (Studi : Institut Teknologi Sepuluh Nopember)' diperoleh hasil bahwa metode *K Nearest Neighbor* melakukan klasifikasi dengan lebih baik dibandingkan Naïve Bayes Classifier dengan tingkat akurasi sebesar 90,26%.

Berdasarkan latar belakang di atas, maka penelitian ini diberi judul 'Analisis Sentimen Pada Ulasan Aplikasi Gojek Menggunakan Metode *Support Vector Machine* (SVM) dan *K Nearest Neighbor* (KNN)'. Penelitian ini bertujuan untuk membandingkan tingkat akurasi antara metode *Support Vector Machine* (SVM) dan *K Nearest Neighbor* (KNN).

METODE

Secara umum tahapan dalam penelitian ini yaitu pengumpulan data, analisis data, dan penarikan kesimpulan. Data yang digunakan dalam penelitian ini diperoleh dari ulasan aplikasi Gojek di website Google Playstore dengan cara *web scraping* sebanyak 2.462 ulasan dengan variabel sebagai berikut.

Tabel 1. Variabel Dataset Ulasan Gojek

Variabel	Definisi Operasional Variabel
<i>Name</i>	Nama pemberi review
<i>Date</i>	Tanggal pengguna memberikan ulasan
<i>Rating</i>	Penilaian pengguna aplikasi Gojek dalam bentuk bintang 1,2,3,4, atau 5
<i>Review</i>	Deskripsi isi ulasan pengguna aplikasi Gojek

K Nearest Neighbor

K Nearest Neighbor (KNN) merupakan salah satu metode dalam machine learning yang digunakan dalam proses pengklasifikasian. Prinsip kerja *K Nearest Neighbor* adalah melakukan klasifikasi berdasarkan kedekatan lokasi (jarak) suatu data dengan data yang lain (Eko Prasetyo, 2013). Dekat atau jauhnya lokasi (jarak) bisa dihitung melalui salah satu dari besaran jarak yang telah ditentukan yakni jarak *Euclidean*, jarak *Minkowski*, dan jarak *Mahalanobis* (Deokar & El-Gayar, 2009). Namun dalam penerapannya seringkali digunakan jarak *Euclidean* karena memiliki tingkat akurasi dan juga *productivity* yang tinggi (Goldfarb, 1991). Jarak *Euclidean* adalah besarnya jarak suatu garis lurus yang menghubungkan antar objek. Rumus jarak

Euclidean adalah sebagai berikut (Aggarwal, 2018):

$$\text{Distance}(\bar{X}, \bar{Y}) = \sqrt{\sum_{n=1}^d (x_i - y_i)^2} \quad (1)$$

dengan:

x_i = data testing X

y_i = data testing Y

$\text{Distance}(\bar{X}, \bar{Y})$ = Jarak Euclid data X dan Y

d = dimensi data variabel bebas

Support Vector Machine

Support Vector Machine (SVM) merupakan suatu teknik yang digunakan untuk melakukan prediksi, baik dalam kasus regresi maupun klasifikasi. Prinsip kerja SVM yaitu mencari *hyperplane* yang optimal dengan margin maksimum yang bertindak sebagai batas keputusan, untuk memisahkan dua kelas yang berbeda (Goh & Lee, 2019). Kasus klasifikasi yang secara linier dapat dipisahkan menjadi prinsip dasar cara kerja SVM, namun SVM telah mengalami perkembangan agar dapat bekerja pada problem non-linier yaitu dengan memasukkan konsep kernel pada ruang kerja berdimensi tinggi. SVM memiliki prinsip utama yaitu menemukan *hyperplane* terbaik yang berfungsi memisahkan dua buah kelas pada ruang input. *Hyperplane* tersebut dapat berupa garis pada *two-dimension* dan dapat berupa *flat plane* pada *multipleplane*.

Pencarian lokasi *hyperplane* optimal merupakan inti dari metode SVM. Diasumsikan bahwa terdapat data *learning* dengan data *points* x_i ($i = 1, 2, \dots, m$) memiliki dua kelas $y_i = \pm 1$ yaitu kelas positif (+1) dan negatif (-1) sehingga akan diperoleh *decision function* berikut.

$$f(x) = \text{sign}(w \cdot x + b) \quad (2)$$

dengan (\cdot) merupakan skalar sehingga

$$w \cdot x = w^T x$$

Margin maksimum dapat diperoleh dengan cara memaksimalkan nilai jarak antara *hyperplane* dan titik terdekatnya yaitu $\frac{1}{\|w\|}$. Hal tersebut

dirumuskan sebagai *Quadratic Programming (QP) Problem* dengan mencari titik minimal seperti pada persamaan berikut.

$$\min \tau(w) = \frac{1}{2} \|w\|^2 \quad (3)$$

Sedangkan subjek *constraint* atau kendala persamaannya adalah sebagai berikut.

$$y_i(w \cdot x_i + b) \geq 1 \quad \forall i \quad (4)$$

Persamaan di atas merupakan permasalahan optimisasi kendala dimana diminimalkan fungsi objek pada persamaan (3) dengan kendala pada persamaan (4). Permasalahan di atas dapat direduksi dengan menggunakan fungsi *Lagrange* yang terdiri dari jumlahan fungsi objektif dan m kendala dikalikan dengan pengganda *Lagrange* seperti berikut (Campbell & Yiming, 2011).

$$L(w, b) = \frac{1}{2} (w \cdot w) - \sum_{i=1}^m \alpha_i (y_i(w \cdot x_i + b) - 1) \quad (5)$$

Dimana α_i merupakan *Lagrange Multipliers*, dan nilai $\alpha_i \geq 0$.

Fungsi kernel digunakan untuk memetakan dimensi awal (dimensi yang lebih rendah) himpunan data ke dimensi baru (dimensi yang relatif lebih tinggi). Macam-macam fungsi kernel (Aggarwal, 2018):

1. Kernel Linear

$$K(x, y) = x \cdot y$$

2. Kernel Gaussian Radial Basic Function (RBF)

$$K(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right)$$

3. Kernel Polynomial

$$K(x_i, x_j) = ((x_i \cdot x_j) + c)^d$$

dengan x_i dan x_j merupakan pasangan dua data training. Parameter $\sigma, c, d > 0$ merupakan konstanta.

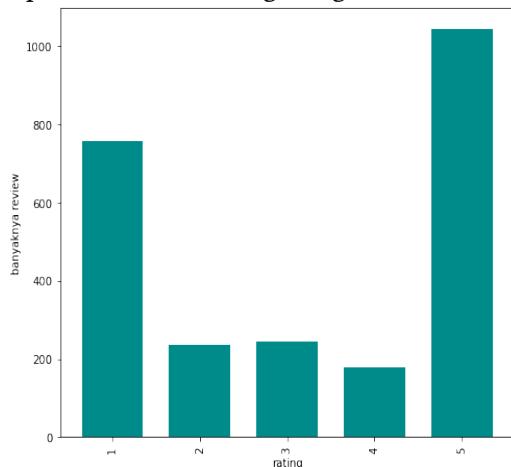
Langkah-langkah analisis data dalam penelitian ini adalah sebagai berikut. (1) Pengumpulan data menggunakan cara *web scraping*. (2) *Text preprocessing* untuk membersihkan data yang meliputi *case folding*,

filtering, stemming, dan tokenizing. (3) Pembobotan kata menggunakan *term frequency – inverse document frequency*. (4) Pelabelan data menggunakan lexicon dari library python. (5) Pembagian data latih dan data uji (6) Penentuan nilai K terbaik untuk metode K Nearest Neighbor dan parameter terbaik untuk kernel-kernel pada metode Support Vector Machine. (7) Pengklasifikasian sentimen menggunakan metode Support Vector Machine (SVM) dan K Nearest Neighbor (KNN). (8) Perbandingan tingkat akurasi antara metode Support Vector Machine (SVM) dan K Nearest Neighbor (KNN). (9) Visualisasi data menggunakan *word cloud*.

HASIL DAN PEMBAHASAN

Analisis Deskriptif

Dari 2.462 sampel ulasan data yang telah diperoleh menggunakan teknik web scraping diperoleh sebaran rating sebagai berikut.



Gambar 1. Sampel Rating Ulasan Gojek

Berdasarkan Gambar 1 dapat diperoleh informasi bahwa rating yang diberikan sangat variatif, rating dengan nilai 5 memperoleh frekuensi paling banyak yaitu 1044, namun rating dengan nilai 1 juga memperoleh frekuensi yang tinggi yaitu 756. Hal tersebut membuat penelitian analisis sentimen begitu diperlukan untuk mengevaluasi ulasan para pengguna aplikasi supaya perusahaan dapat mengambil keputusan terbaik dalam menjaga dan meningkatkan mutu pelayanannya.

K Nearest Neighbor

Metode pertama yang digunakan untuk mengklasifikasi data ulasan pengguna Aplikasi Gojek adalah KNN. Hal terpenting dalam melakukan klasifikasi menggunakan metode

KNN adalah penentuan nilai K terbaik untuk membuat model klasifikasi. Dalam penelitian ini akan digunakan cara *trial and error* pada nilai $K=1, 2, 3, \dots, 40$. Diperoleh beberapa nilai K terbaik dengan tingkat akurasi sebagai berikut.

Tabel 2. Penentuan nilai K terbaik KNN

Nilai K Akurasi Presisi Recall

$K = 14$	81,65%	83,36%	92,69%
$K = 16$	81,33%	82,75%	93,15%
$K = 22$	82,14%	82,28%	95,43%
$K = 26$	81,65%	81,67%	95,66%

Berdasarkan Tabel 2 di atas, dapat diinformasikan bahwa $K=22$ memiliki tingkat akurasi yang paling tinggi, yaitu 82,14%. Dengan demikian model klasifikasi K Nearest Neighbor dengan $K=22$ ini akan dibandingkan dengan hasil pengujian tingkat akurasi pada metode Support Vector Machine pada pembahasan berikutnya.

Support Vector Machine

Metode kedua yang digunakan untuk mengklasifikasi data ulasan pengguna Aplikasi Gojek adalah SVM, dengan tiga fungsi kernel, yaitu fungsi kernel Linear, fungsi kernel RBF, dan fungsi kernel Polynomial. Dengan cara *trial and error* diperoleh hasil parameter terbaik dari masing-masing kernel sebagai berikut.

Tabel 3. Uji Parameter Kernel Pada SVM

Jenis kernel	Akurasi	Presisi	Recall
Linear ($C=1$)	87,98%	88,55%	95,43%
RBF	86,85%	86,65%	96,43%
Polinomial	84,25%	88,31%	89,72%

Berdasarkan Tabel 3 di atas, dapat diinformasikan bahwa kernel linear dengan parameter $C=1$ memiliki tingkat akurasi tertinggi, yaitu 87,98%. Dengan demikian model klasifikasi Support Vector Machine dengan kernel linear dimana $C=1$ ini akan dibandingkan dengan tingkat akurasi metode klasifikasi K Nearest Neighbor pada pembahasan berikutnya.

Perbandingan KNN dan SVM

Setelah didapatkan hasil tingkat akurasi pada kedua metode, maka langkah selanjutnya adalah membandingkannya. Berikut merupakan perbandingan antara metode Support Vector Machine dan metode K Nearest Neighbor.

Tabel 4. Perbandingan Akurasi SVM dan KNN

Metode	Akurasi
Support Vector Machine (C=1)	87,98%
K Nearest Neighbor (K=22)	82,14%

Berdasarkan Tabel 4 dapat disimpulkan bahwa metode SVM memiliki kinerja lebih baik dibandingkan metode KNN dalam melakukan klasifikasi analisis sentimen pada ulasan pengguna aplikasi Gojek di Google Playstore.

Visualisasi Data dengan Word Cloud

Visualisasi data teks menggunakan *word cloud* digunakan untuk mengetahui kata-kata yang paling sering muncul dalam dokumen. *Word Cloud* memvisualisasi kata-kata berdasarkan kategori sentimennya, sehingga dapat diketahui kata-kata yang sering digunakan pada masing-masing sentimen. Semakin besar ukuran font maka berarti semakin besar frekuensi kemunculan kata tersebut. Berikut merupakan *word cloud* pada analisis sentimen ulasan pengguna aplikasi gojek.



Gambar 2. Word Cloud Sentimen Positif



Gambar 3. Word Cloud Sentimen Negatif

PENUTUP

Dari 2.462 sampel ulasan aplikasi Gojek yang diambil secara acak dari Google Playstore diperoleh hasil bahwa 1.751 memiliki sentimen positif dan 711 memiliki sentimen negatif. Berdasarkan kata-kata yang sering muncul dalam sentimen negatif, dapat ditarik kesimpulan bahwa para pengguna aplikasi yang memberikan sentimen negatif memiliki keluhan yang berhubungan dengan akurasi lokasi peta, notifikasi akun yang mengalami error, dan kebijakan biaya admin paylater yang dianggap terlalu mahal.

Hasil pengujian pada Metode SVM diperoleh hasil bahwa kernel *linear* dengan parameter $C=1$ memperoleh nilai terbaik dengan akurasi, presisi, dan recall berturut-turut sebesar 87,98%, 88,55%, dan 95,43%. Hasil pengujian pada Metode KNN dengan nilai $K=22$ sebagai nilai K terbaiknya memperoleh nilai akurasi, presisi, dan recall berturut-turut sebesar 82,14%, 82,28%, dan 95,43%. Dengan demikian dapat disimpulkan bahwa metode *Support Vector Machine (SVM)* melakukan klasifikasi secara lebih baik dibandingkan *K Nearest Neighbor (KNN)* pada ulasan pengguna aplikasi Gojek di Google Playstore.

DAFTAR PUSTAKA

- Aggarwal, C.C. 2018. *Machine Learning for Text*. New York: Springer International Publishing.

Campbell, C. & Yiming, Y. 2011. *Learning with Support Vector Machines*. London: Morgan&Claypool Publishers

Deokar, A. V. & El-Gayar, O.F. 2009. *Enabling distributed model management using semantic web technologies*. Proceedings of the 42nd Annual Hawaii International Conference on System Sciences, HICSS. Hawaii 5-8 January 2009

Eko Prasetyo. 2013. *Data Mining : Konsep Dan Aplikasi Menggunakan Matlab*. Yogyakarta: CV Andi Offset

Goh, R.Y. & Lee, L.S. 2019. Credit Scoring: A Review on Support Vector Machines and Metaheuristic Approaches. *Hindawi Limited*. 30(1): 11-19

Hidayat, A.N. 2015. Analisis Sentimen Terhadap Wacana Politik Pada Media Masa Online Menggunakan Algoritma Support Vector Machine Dan Naive Bayes. *Jurnal Elektronik Sistem Informasi Dan Komputer (Jesik)*, 1(1): 1-7

Kurniawan, T. 2017. Implementasi Text Mining Pada Analisis Sentimen Pengguna Twitter Terhadap Media Mainstream Menggunakan Naïve Bayes Classifier Dan Support Vector Machine Media Mainstream Menggunakan Naïve Machine. *ITS Journal*. 23(1): 1-6.

Lestari, D.R.D. 2019. Perbandingan Klasifikasi Beasiswa Toyota Astra Menggunakan K-Nearest Neighbor Classifier dan Nive Bayes Sebagai

Penentu Metode Klasifikasi Pada Sistem Pendukung Keputusan Penerimaan Beasiswa Toyota Astra (Studi : Institut Teknologi Sepuluh Nopember). *Journal of Chemical Information and Modeling*. 53(9): 1689–1699

Liu, B. 2016. Opinion Mining. *Encyclopedia of Database Systems*. Chicago: University of Illinois at Chicago

Wayan, N. & Saraswati, S. 2013. *Naïve bayes classifier dan support vector machines untuk sentiment analysis*. Seminar Nasional Sistem Informasi Indonesia. Bali 2-4 Desember 2013