Inserch
Information System Research Journal

E-ISSN: 2775-4669
Volume 2, Nomor 1, Februari 2022

Analisis Sentimen pada Sosial Media Twitter terhadap MRT Jakarta Menggunakan Machine Learning

Dina Agustina[1], Fitri Rahmah[2]
Matematika, FMIPA, Universitas Negeri Padang[1]
Jln. Prof. Dr. Hamka Air Tawar Padang[1]
Ilmu Ekonomi, Ekonomi,Universitas Andalas, Padang[2]
dinagustina@fmipa.unp.ac.id[1], fitrirahmah26@yahoo.com[2]

Diterima: 22 Dec 2021 | Direvisi: 19 Jan 2022
Disetujui: 20 Jan 2022 | Dipublikasi: 16 Mar 2022

Abstrak

Moda raya terpadu (MRT) merupakan moda transportasi terbaru yang ada di ibukota untuk mengatasi kemacetan. Jalur MRT terdiri dari 16 rute yang berada di area-area strategis seperti pusat perkantoran, pebelanjaan serta daerah pemukiman. Untuk melihat opini masyarakat Jakarta terkait dengan baru beroperasinya MRT maka dilakukan analisis opini masyarakat Jakarta berdasarkan laman social media masyarakat khususnya *twitter*. Opini ini digunakan untuk melihat analisis sentimen (positif, netral dan negatif) dari masyarakat kota Jakarta terkait keberadaan MRT yang baru beroperasi. Hasil sentimen ini dapat digunakan untuk melihat pandangan masyarakat Jakarta terkait layanan moda transportasi darat baru yang disediakan oleh pemerintah. Metode yang digunakan pada penelitian ini adalah algoritma *machine learning* klasifikasi, yaitu Naïve Bayes. Dataset terdiri dari 2268 *tweet* masyarakat Jakarta. Dilakukan *exploratory data analysis* (EDA) untuk melihat sentimen masyarakat. Diperoleh bahwa persentase sentimen positif (48,8%), sentimen negatif (22.4%) dan netral (28.8%). Dataset yang sudah *clean* dibagi menjadi data *training* dan *testing*. Pada data *training* diaplikasikan algoritma *machine learning* untuk memperoleh model klasifikasi dan menentukan nilai akurasi. Diperoleh model analisis *text mining* dengan Naïve Bayes memiliki akurasi sebesar 76,21%.

Kata Kunci: Machine learning, Naïve Bayer, MRT

Abstract

The integrated highway (MRT) is the newest mode of transportation in the capital to overcome congestion. The MRT line consists of 16 routes located in strategic areas such as office centers, shopping and residential areas. To see the opinion of the people of Jakarta related to the new operation of the MRT, the opinion of the people of Jakarta is analyzed based on the community's social media pages, especially Twitter. This opinion is used to look at the sentiment analysis (positive, neutral and negative) of the people of Jakarta regarding the existence of the new MRT operating. The results of this sentiment can be used to see the views of the people of Jakarta regarding the new land transportation services provided by the government. The method used in this research is a classification machine learning algorithm, namely Naïve Bayes. The dataset consists of 2268 tweets from the people of Jakarta. An exploratory data analysis (EDA) was conducted to see public sentiment. It was found that the percentage of positive sentiment (48.8%), negative sentiment (22.4%) and neutral (28.8%). The clean dataset is divided into training and testing data. In the training data, machine learning algorithms are applied to obtain a classification model and determine the accuracy value. The analytical model using Naïve Bayes has an accuracy of 76.21%.

Keyword: Machine learning, Naïve Bayer, MRT

I. PENDAHULUAN

Kemacetan merupakan permasalahan yang terjadi di kota-kota besar, salah satunya yaitu Jakarta, ibukota Indonesia. Bertambahnya jumlah penduduk (urbanisasi) di Jakarta mendorong masyarakatnya menggunakan berbagai

transportasi untuk mobilisasi. Pada umumnya masyarakat kota cenderung menggunakan kendaraan pribadi yang menyebabkan volume kendaraan semakin banyak yang memicu terjadinya kemacetan. Penyediaan layanan transportasi yang nyaman, mudah, murah serta terjangkau menjadi salah satu solusi yang ditawarkan oleh pemerintah untuk mengurangi volume kendaraan pribadi sehingga angka kemacetan dapat ditekan. Pemerintah Jakarta membangun moda transportasi baru yaitu Moda Raya Terpadu (MRT). Jalur MRT terdiri dari 16 rute yang berada di area-area strategis seperti pusat perkantoran, pebelanjaan serta daerah pemukiman. MRT pertama kali beroperasi pada tanggal 24 Maret 2019.

Untuk mengetahui apakah keberadaan transportasi baru MRT menjadi salah satu alternatif transportasi yang dapat menarik minat pengguna transportasi pribadi ataupun lainnya yang dapat mengurangi angka kemacetan maka perlu dilakukan analisis terkait opini pengguna layanan. Analisis yang dilakukan terkait sentimen masyarakat dengan keberadaan MRT. Berdasarkan lembaga riset media social Semiocast yang berada di Paris, Negara Indonesia berada pada lima besar dengan jumlah pemilik akun *twitter* terbesar di dunia dan menempati urutan ketiga Negara yang paling aktif menggunakan *twitter* perhari [1]. Oleh sebab itu, data yang digunakan sebagai analisis sentiment adalah opini masyarakat Jakarta dari aktifitas media sosial, *twitter*.

Opinion mining atau bisa disebut juga analisis sentimen merupakan suatu proses deteksi polaritas dan ekstraksi fitur (variabel) dalam bentuk klasifikasi dengan kategori positif dan negatif [2]. Ini dapat dijadikan sebagai analisis sentimen masyarakat Jakarta terhadap baru beroperasinya MRT dengan menggunakan *twitter*. Informasi dari data *twitter* diolah dengan mengkategorikan opini tersebut ke kelas sentiment positif atau negative. [3] melakukan penelitian analisis sentimen berdasarkan *tweet* menggunakan *K-neighbour classifier*. Metode seperti Naïve Bayes Classification, Maximum Entropy, ataupun *Support Vector Machine* digunakan sebagai analisis sentimen terhadap *twitter* [4][5]. Metode Naïve Bayes merupakan metode yang memiliki tingkat akurasi yang paling tinggi terkait dengan analisis dokumen tekstual [6].

Berdasarkan latar belakang tersebut dilakukan penelitian mengenai analisis sentimen masyarakat Jakarta terhadap baru beroperasinya MRT melalui social media *twitter* menggunakan algortima *machine learning* (ML) klasifikasi, Naïve Bayes.

II. TINJAUAN PUSTAKA

A. Twitter dan Analisis Sentimen

Microblogging, Twitter, telah menjadi alat komunikasi yang sangat populer di kalangan pengguna Internet dalam beberapa tahun terakhir [7]. Twitter memberikan layanan kepada penggunanya untuk mengirim dan membaca pesan berbasis teks hingga 280 karakter yang dikenal dengan sebutan kicauan (*tweet*).

Adapun teknik *opinion mining* (analisis sentimen) pada ML yaitu Naïve Bayes, SVM dan *multilayer perception (*MLP) [8]. Teknik ini digunakan untuk menambang data dan mengolahnya menjadi informasi berguna (*insight*). Informasi tersebut diklasifikasikan pada kelompok sentiment positif, netral atau negatif.

B. Machine Learning

Machine learning merupakan salah satu cabang dari *artificial intelligence* (AI). ML pertama kali diperkenalkan olej ilmuwan matematika seperti Adrien Marie Legendre, Thomas Bayes dan Andrey Markov pada tahun 1920. Secara umum, ML dikelompokan ke menjadi 2 jenis yaitu ML *supervised* dan *unsupervised* [9].

Supervised learning adalah teknik pembelajaran yang menentukan pola hubungan antara variabel input dengan output (label) dari data training yang diberikan sedangkan *unsupervised learning* adalah teknik yang bertugas untuk menyimpulkan sebuah fungsi untuk menggambarkan struktur tersembunyi dari *data training* yang tidak berlabel. Proses kerja ML berdasarkan dataset yang di *split* menjadi data *training* dan data *testing*. Kemudian data *traning* dimasukan dalam algoritma ML untuk menghasilkan sebuah model prediksi. Data testing digunakan untuk menentukan hasil prediksi berdasarkan model prediksi. Cerdas atau tidaknya suatu ML bergantung pada kualitas data training, pemilihan fitur serta algoritma yang digunakan. Pada penelitian ini digunakan ML *supervised* dengan metode Naïve Bayes klasifikasi.

C. Metode Naïve Bayes

Naive Bayes klasifikasi adalah model *machine learning* probabilistik yang digunakan

untuk klasifikasi. Naive Bayes mengasumsikan bahwa keberadaan fitur tertentu di kelas tidak terkait dengan keberadaan fitur lainnya.

Teorema Bayes memberikan cara menghitung probabilitas posterior $P(c|x)$ dari P $(c), P(x)$ dan $P(x|c)$ [10]. Adapun formula teorema Bayes yaitu:

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)}$$

$$P(c|X) = P(x_1|c) \times P(x_2|c) \times \ldots \times P(x_n|c) \times P(c)$$

Dimana:

1. $P(c|x)$ adalah probabilitas posterior kelas (c, target) dengan prediktor (x, atribut).
2. $P(c)$ adalah probabilitas prior kelas.
3. $P(x|c)$ adalah likelihood yang merupakan probabilitas kelas yang diberikan.
4. $P(x)$ adalah probabilitas prior predictor.

Naive Bayes *classifier* mengasumsikan bahwa efek dari nilai *predictor* atau nilai fitur (x) pada *given class* (c) adalah independen terhadap nilai-nilai dari prediktor atau fitur lainnya.

Langkah-langkah model Naïve Bayes untuk analisis sentimen adalah

1. *Dataset* yang berupa *tweet* dibagi menjadi data training dan testing dengan perbandingan (90%:10%).
2. Setiap *tweet* diberi label atau dikategorikan ke dalam positif, netral atau negatif.
3. Kata pada *tweet* didefinisikan sebagai fitur (variabel) yang saling independen, kemudian diidentifikasi daftar kata yang muncul.
4. Probabilitas kemunculan setiap kata dihitung kemudian dikategorikan pada kelompok positif, kelompok netral dan kelompok negative atau dengan menggunakan formula *[P(x)=predictor prior probability]*
5. Dihitung probabilitas kemunculan positif, probabilitas kemunculan netraldan probabilitas kemunculan negatif atau dengan menggunkan formula *[P(c)=class prior probability]*
6. Menghitung *likelihood [P(x|c)]* dan *posterior probability [P(c|x)]* untuk masing-masing kata atau fitur pada dataset.

Adapun kelebihan dan kelemah metode Naïve Bayes yaitu:

a. Kelebihan

1. Mudah dan cepat untuk memprediksi kelas data *testing*. Naive Bayes dapat digunakan dalam prediksi multikelas.
2. Jika asumsi independensi berlaku, Naive Bayes berperforma lebih baik dibandingkan dengan model lain seperti regresi logistik dan memerlukan lebih sedikit data pelatihan (*training*).
3. Berkinerja baik pada variabel input kategori dibandingkan dengan variabel numerik. Untuk variabel numerik, diasumsikan berdistribusi normal (kurva lonceng, yang merupakan asumsi kuat).

b. Kelemahan

1. Jika variabel kategorikal memiliki kategori (dalam kumpulan data test), yang tidak diamati dalam kumpulan data pelatihan, maka model akan menetapkan probabilitas 0 (nol) dan tidak akan dapat membuat prediksi. Ini sering disebut sebagai "Frekuensi Nol". Untuk mengatasinya, kita bisa menggunakan teknik *smoothing*. Salah satu teknik *smoothing* yang paling sederhana disebut estimasi *Laplace*.
2. Di sisi lain bayes naif juga dikenal sebagai penduga yang buruk, sehingga keluaran probabilitas dari prediksi probabilitas tidak dianggap terlalu serius.
3. Limitasi dari Naive Bayes adalah asumsi prediktor independen. Dalam kehidupan nyata, hampir tidak mungkin kita mendapatkan seperangkat prediktor yang sepenuhnya independen.

Aplikasi algoritma Naïve Bayes dalam kehidupan sehari-hari yaitu prediksi real time, prediksi multi kelas, klasifikasi teks / pemfilteran spam/analisis sentimen dan sistem rekomendasi

D. Model Akurasi

Perhitungan Performansi algoritma Naïve Bayes Classifier adalah dengan menggunakan *confusion matrix*. Performansi dihitung dari hasil klasifikasi. Berikut table perhitungan *confusion matrix.*

TABEL I. CONFUSION MATRIX

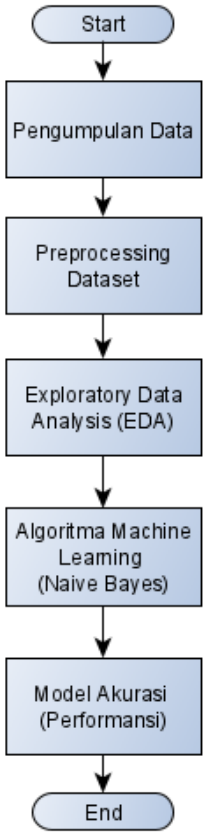
		Aktual	
		Positif	**Negatif**
Prediksi	**Positif**	True Positive (TP)	False Positive (FP)
	Negatif	False Negatif (FN)	True Negatif (TN)

Formula untuk menghitung akurasi dari algoritma Naïve Bayes:

$$Accuracy = \frac{TP + TN}{n}$$

III. METODE PENELITIAN

Metode pengolahan data yang digunakan merupakan metode klasifikasi *machine learning* yaitu algoritma Naïve Bayes dengan menggunakan *software* Python. Data yang digunakan merupakan data yang diperoleh melalui aktifitas social media yaitu *twitter* masyarakat Jakarta. Dataset terdiri dari 2268 *tweet*. Berikut *flowchart* dari penelitian ini.



Gambar 1. *Flowchat* Penelitian

Uraian dari setiap tahapan pada *flowchart* adalah sebagai berikut:

1. Pengumpulan Data *Tweet*

   Data twitter yang dikumpulkan berasal dari *tweet* masyarakat Jakarta pada saat MRT Beroperasi. Jumlah dataset yang terkumpul adalah 2268.

2. *Preprocessing* Data

   Tahapan *preprocessing* data terdiri dari

   a. Pelabelan *tweet* ke dalam kategori positif, netral dan negatif dengan memanfaatkan konsep Naïve Bayes.

   b. *Cleaning* data dengan menghapus *stopword* (data yang sering muncul dan tidak ada makna), *stemming* data (penguraian data menjadi kata dasar), *remove username*, url, *hashtags* dan *punctuation*.

   c. Setelah data *clean*, data dibagi menjadi data *training* dan data *testing* dengan komposisi (90%:10%).

   Contoh hasil data *clean* dapat dilihat pada tabel II berikut.

TABEL II. HASIL DATA CLEANING

Tweets	Label	Clean
b'@zahirarra ayo jalan2 naik mrt'	Positif	ayo jalan2 mr
b'yeay! (@ stasiun mrt bundaran hi in jakarta,...	Netral	yeay stasiun mrt bundar jakarta dki jakarta
b'ngeliat rute kantor sekarang. naik motor salah. naik mrt salah…	Negatif	ngeliat rute kantor motor salah mrt salah nmem…

3. *Exploratory Data Analysis* (EDA)

   Data yang sudah clean dilakukan EDA untuk untuk memperoleh *insight* data. Adapun proses yang dilakukan adalah sebagai berikut:

   a. Mengubah label menjadi label kategori (positif, netral, negatif diganti dengan 2,1,0). Contoh dari label yang sudah dirubah bisa dilihat pada Tabel III berikut.

TABEL III DATA YANG SUDAH DILABELI

Tweets	Label	Clean
b'@zahirarra ayo jalan2 naik mrt'	2	ayo jalan2 mr
b'yeay! (@ stasiun mrt bundaran hi in jakarta,...	1	yeay stasiun mrt bundar jakarta dki jakarta
b'ngeliat rute kantor sekarang. naik motor salah. naik mrt salah…	0	ngeliat rute kantor motor salah mrt salah nmem…

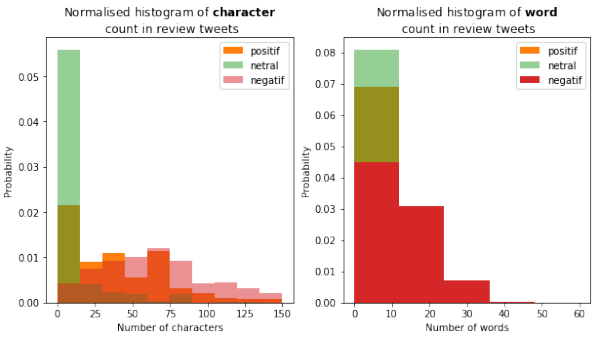
   b. Dibuat tabel list kata yang sering muncul pada tiap sentimen (positif, netral, dan negatif), kemudian tabel digabungkan.

Hasil *list* kata dapat dilihat pada gambar 2 berikut.

	positive	netral	negative	total
mrt	908	554	437	1899
lrt	28	444	39	511
mrt lrt	7	442	10	459
train	3	442	3	448
train mrt	1	441	0	442

Gambar 2. List 5 Kata Teratas Setiap Sentimen

c. Divisualisaikan kata dan panjang kata tiap sentimen (positif, netral, dan negatif)



Gambar 3. Sebaran kata dan panjang kata

4. Algoritma Naïve Bayes

Dilakukan analisis dengan menggunakan metode Naïve Bayes pada data *training*. Data berbentuk teks di-*convert* menjadi data numerik dengan menggunakan library *CountVectorizer* dan *TfidfTransformer*.Selanjutnya diaplikasikan Algoritma ML Naïve Bayes klsifikasi dengan menggunakan library *sklearn.naive_bayes* dengan mengimpor modul *BernoulliNB* sehingga diperoleh model predictor (klasifikasi) Naïve Bayes.

5. Model Akurasi

Performansi atau keakuratan model Naïve Bayes dihitung dari menginputkan data *testing* pada model predictor yang telah diperoleh. Dari proses tersebut diperoleh hasil klasifikasi. Selanjutnya dicari nilai *confusion matrix*. Pada penelitian ini *confusion matrix* dihitung dengan menggunakan library *sklearn.metrics* dengan mengimport modul *accuracy_score, confusion_matrix.*

IV. HASIL DAN PEMBAHASAN

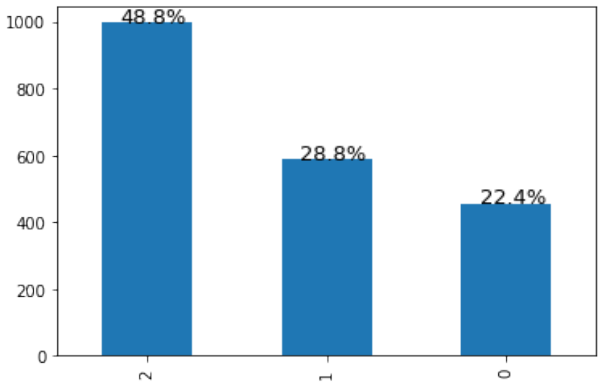
Berikut akan dibahas hasil dari analisis *text mining* dan nilai akurasi dari metode Naïve Bayes Klasifikasi.

A. Analisis Sentimen

Berdasarkan hasil analisis text mining atau EDA yang dilakukan pada data *tweet* yang telah dilakukan preprocessing diperoleh hasil analisis sentiment masyarakat Jakarta terkait baru beroperasinya MRT yaitu:

1. Sentimen positif masyarakat terkait keberadaan MRT sebesar 48,8%. Dari hasil ini berarti, sebagian besar masyarakat Jakarta menyambut dengan baik moda transportasi baru yang ditawarkan oleh pemerintah.
2. Sentimen netral sebesar 28,8%, artinya beberapa masyarakat menunjukkan sikap biasa saja dengan adanya MRT.
3. Sebesar 22.4% menunjukkan sentimen negatif dengan keberadaan MRT.

Berikut divisualisasikan sebaran dari analisis sentiment berdasarkan kategori positif, netral dan Negatif yang dapat dilihat pada gambar 4.



Gambar 4. Distribusi Sentimen

Dari hasil analisis sentiment ini dapat kita lihat, kecenderungan masyarakat Jakarta menyambut baik moda transportasi MRT yang baru beroperasi. Berdasarkan analisis opini masyarakat tersebut, diharapkan MRT bisa menjadi salah satu solusi untuk menguraikan kemacetan atau mengurangi volume kendaraan pribadi di jalanan ibukota.

B. Nilai Akurasi

Keakuratan model dianalisis menggunakan *confusion matrix*. Dari hasil penginputan data *testing* ke model prediksi diperoleh nilai akurasi Naïve Bayes sebesar 76,21 %. Model akurasi Naïve Bayes yang diperoleh belum terlalu

bagus. Hal ini dikarenakan *preprocessing* data pada data *cleaning* untuk *text mining* belum optimal. Pada proses *stopwords* dengan bahasa Indonesia sangat terbatas sehingga diperlukan untuk menginput sendiri *stopwords* Indonesia.

V. KESIMPULAN

Dari hasil EDA dan model analisis *text mining* diperoleh bahwa;

a. Persentase sentimen positif (48,8%) lebih tinggi dari pada sentimen negatif (22.4%) dan netral (28.8%) masyarakat.

b. Model analisis text mining dengan Naïve Bayes memiliki nilai akurasi 76,21%.

Saran untuk penelitiaan lebih lanjut dari analisis sentiment dengan data *twitter* yaitu:

a. Dibutuhkan kamus khusus bahasa Indonesia untuk *stopwords* pada proses data *cleaning* sehingga hasil analisis sentiment lebih akurat.

b. Data hasil sentiment analisis dapat digunakan pihak terkait untuk peningkatan layanan dan kenyamanan bagi pengguna transportasi.

REFERENSI

[1] kominfo.go.id/content/detail/3415/kominfo-pengguna-internet-di-indonesia-63-juta-orang/0/berita_satker

[2] Alsaeedi, A. & Khan, M. Z., 2019. "A Study on Sentiment Analysis Techniques of Twitter Data". International Journal of Advanced Computer Science and Applications (IJACSA), 10(2), pp. 361-374.

[3] Anurag P. Jain, Vijay D. Katkar, "Sentimen Analysis of Twitter Data Using Data Mining", International Conference on Information Processing (ICIP), IEEE, 13 Juni 2016.

[4] M. Y. Nur and D. D. Santika, "Analisis Sentimen pada Dokumen Berbahasa Indonesia dengan Pendekatan Support Vector Machine," in Konferensi Nasional Sistem dan Informatika, Bali, 2011

[5] A. Go, R. Bhayani, and L. Huang, "Twitter Sentiment Classification using Distant Supervision," Stanford, Project Report CS224N, 2009.

[6] Hairani, G. S. Nugraha, M. N. Abdillah, M. Innuddin, "Komparasi Akurasi Metode Correlated Naive Bayes Classifier dan Naive Bayes Classifier untuk Diagnosis Penyakit Diabetes", Jurnal Nasional Informatika dan Teknologi Jaringan, Vol 3 No. 1, September 2018.

[7] Liang, Po-Wei & Dai, Bi-Ru. (2013). "Opinion Mining on Social Media Data". 2. 91-96. 10.1109/MDM.2013.73.

[8] R, Mewari. A. Singh, A. Srivastava. 2015. "Opinion Mining Techniques on Social Media Data ". International Journal of Computer Applications (0975 – 8887) Volume 118 – No. 6.

[9] G. James, D. Witten, T. Hastie, R. Tibshirani. "An Introduction to Statistical Learning with Application in R". 2017. Springer

[10] B. Lantz. "Machine Learning with R". 2013. Packt Publishing Ltd.