Jurnal Elektronik Ilmu Komputer Udayana
Volume 13, No 2. November 2024

p-ISSN: 2301-5373
e-ISSN: 2654-5101

Penerapan *Machine Learning* Dalam Analisis Sentimen dan Pemodelan Topik Data Opini Kendaraan Listrik

Putu Ayu Novia Aryanti[a1], I Made Widhi Wirawan[a2], Anak Agung Istri Ngurah Eka Karyawati[a3], I Wayan Supriana[a4]

[a]Program Studi Informatika, Universitas Udayana
Bali, Indonesia
[1]niputunovia@email.com
[2]made_widhi@unud.ac.id
[2]eka.karyawati@unud.ac.id
[2]wayan.supriana@unud.ac.id

Abstract

Regulation of the Minister of Industry Number 6 of 2023 concerning Guidelines for Providing Government Assistance for the Purchase of Two-Wheeled Battery-Based Electric Motorized Vehicles is one of the government policies aimed at encouraging the growth of the domestic electric vehicle ecosystem. This study aims to analyze public opinion related to electric vehicles on social media Twitter and YouTube. Using the Support Vector Machine (SVM) method, this study classifies public sentiment into positive and negative, and uses Latent Dirichlet Allocation (LDA) to modelling topics in each sentiment. The results show that of the 492 public opinion test data, 244 data are labeled positive and 248 data are labeled negative. By using the parameters of polynomial kernel, C=1.0, and degree=2, the SVM model achieved accuracy, precision, recall, and F1-score of 86%. In addition, the best model for positive sentiment is the model with 3 topics resulting in a coherence value of 0.4437. The best model for negative sentiment is the model with 4 topics resulting in a coherence value of 0.47698. Data with positive sentiment discusses a lot about public support and interest in electric vehicles. Data with negative sentiment discusses a lot about electric vehicle subsidies that are less targeted.

Keywords: *sentiment analysis, topic modelling, electric vehicles, support vector machine, latent dirichlet allocation*

1. Pendahuluan

Kendaraan listrik merupakan alat transportasi bertenaga motor listrik dengan penyimpanan energi listriknya di dalam baterai. Kendaraan listrik dapat meliputi kendaraan beroda empat seperti mobil, maupun kendaraan beroda dua seperti sepeda motor, sepeda, dan skuter. Kendaraan listrik digadang-gadang sebagai kendaraan ramah lingkungan karena tidak menghasilkan emisi yang berbahaya. [1] Sejarah awal kemunculan kendaraan listrik di Indonesia dimulai pada tahun 2012. Namun, kemajuan kendaraan listrik sempat mengalami tantangan. [2] Berdasarkan survei Charta Politika tahun 2022, dari 1220 responden dari seluruh masyarakat Indonesia 61% menyatakan kurang tertarik dengan kendaraan listrik. Keengganan tersebut dipicu oleh beberapa faktor seperti keraguaan responden terkait teknologi yang digunakan, mahalnya harga kendaraan baru, dan kesulitan dalam menemukan penjual kendaraan listrik ataupun stasiun pengisian bahan bakar listrik di daerah tempat tinggal responden. [3]

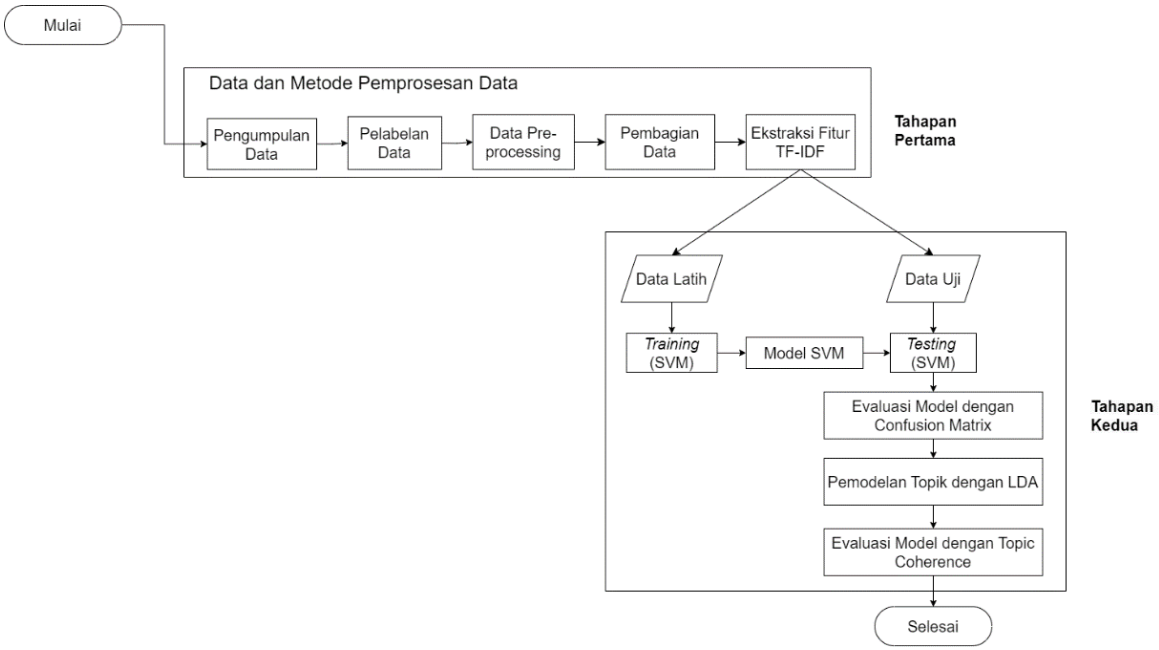
Dalam mengatasi kurangnya minta masyarakat tentang kendaraan listrik, Pemerintah Republik Indonesia mencanakan kebijakan untuk mendorong pertumbuhan ekosistem kendaraan listrik yang dituangkan pada Peraturan Menteri Perindustrian Nomor 6 Tahun 2023 tentang Pedoman Bantuan Pemerintah untuk Pembelian Kendaraan Bermotor Listrik Berbasis Baterai Roda Dua. Program bantuan tersebut berupa potongan harga sebesar Rp 7 juta untuk pembelian satu unit KBL berbasis baterai roda dua dan hanya berlaku untuk setiap Nomor Induk Kependudukan (NIK). [4] Kebijakan

lainnya dalam Peraturan Menteri Keuangan Nomor 38 Tahun 2023 tentang PPN atas Penyerahan Kendaraan Bermotor Listrik Berbasis Baterai Roda Empat dan Kendaraan Baterai Bus Tertentu yang ditanggung Pemerintah tahun anggaran 2023. [5]

Dengan pemikiran ini, penulis tertarik untuk melakukan analisis sentimen dan pemodelan topik opini publik terhadap kendaraan listrik pada media sosial setelah kebijakan tersebut diberlakukan. Masalah analisis sentimen dengan berbagai metode *machine learning* sudah pernah dibahas dalam penelitian – penelitian sebelumnya diantaranya, penelitian [6] membahas mengenai analisis sentimen terhadap opini masyarakat pada sosial media YouTube dengan algoritma Naïve Bayes dengan nilai akurasi sebesar 82% serta polaritas sentimen yang didapat yaitu 82% sentimen negatif dan 18% sentimen positif. Penelitian lainnya [7] membahas mengenai analisis sentimen dan pemodelan topik pada review aplikasi Ruang Guru dengan Support Vector Machine (SVM) dan Latent Dirichlet Allocation (LDA) memperoleh nilai akurasi sebesar 90%. Penelitian terkait analisis sentimen [8] pendapat masyarakat tentang mobil listrik menggunakan metode SVM dan Selection Particle Swarm Optimization dengan tingkat akurasi sebsar 86,07%.

Berdasarkan dari penelitian sebelumnya, pada penelitian ini analisis sentimen dan pemodelan topik kendaraan listrik difokuskan pada obyek berupa opini publik pada Twitter dan YouTube dengan menggunakan salah dua metode *machine learning*. Analisis sentimen dengan menggunakan metode Support Vector Machine (SVM) dengan klasifikasi orientasi sentiment dalam dua jenis yaitu positif dan negatif. Serta pemodelan topik untuk mengungkap topik-topik yang tersembunyi pada hasil analisis sentimen dengan Latent Dirichlet Allocation (LDA).

2. Metode Penelitian



Gambar 1. Diagram Alur Penelitian

Tahapan dari penelitian ditunjukkan pada Gambar 1 yang terbagi menjadi dua tahapan utama. **Tahapan pertama**, berhubungan dengan data dan metode pemrosesan data yang meliputi: proses pengumpulan data opini masyarakat terkait kendaraan listrik dari media sosial Twitter dan YouTube dengan kata kunci “kendaraan listrik”, “mobil listrik”, dan “motor listrik”; pelabelan data oleh ahli; *data preprocessing* untuk membersihkan data yang telah dikumpulkan sehingga siap diproses pada tahapan selanjutnya; ekstraksi fitur TF-IDF; serta pembagian data. **Tahapan kedua** adalah pengembangan metode komputasi berupa pengembangan model analisis sentimen menggunakan *Support Vector Machine* (SVM) yang terdiri dari tahap pelatihan model (*training*), pengujian model (*testing*), dan proses evaluasi model dengan menggunakan *confusion matrix*. Setelah proses analisis sentimen selesai, akan

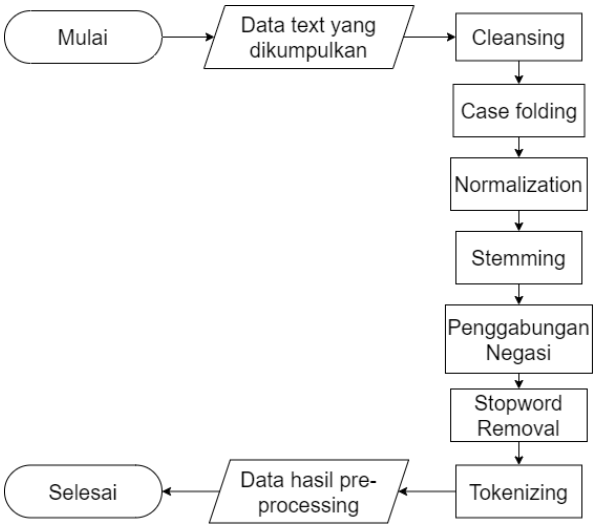
p-ISSN: 2301-5373
e-ISSN: 2654-5101

dilakukan pemodelan topik dengan metode *Latent Dirichlet Allocation* (LDA) pada hasil sentimen positif dan negatif dan evaluasi pemodelan topik dengan menggunakan *topic coherence* guna menentukan topik – topik yang merepresentasikan masing – masing kelas sentimen.

2.1. Pengumpulan Data

Pengumpulan data opini masyarakat tentang kendaraan listrik menggunakan cara *data crawling* yang merupakan metode mengekstrak data secara otomatis dari sosial media dengan Graph API (*Application Programming Interface*). Pengumpulan data dilakukan pada sosial media YouTube dan Twitter dimulai pada tanggal 21 Maret 2023 hingga 30 Juni 2023 setelah ditetapkannya kebijakan baru pemerintah terkait kendaraan listrik, dengan jumlah data yang diambil sekitar 2492 opini dengan hanya data berbahasa Indonesia yang mengandung kata kunci “kendaraan listrik”, “mobil listrik”, dan “motor listrik”. Kemudian data dilabeli secara manual ke dalam sentimen positif atau negatif dengan bantuan ahli bahasa yang merupakan salah satu dosen Program Studi Sastra Indonesia Fakultas Ilmu Budaya Universitas Udayana. Dari hasil pelabel diperoleh data seimbang pada masing – masing sentimen dengan jumlah data untuk setiap label adalah 1246 data.

2.2. Pemrosesan Data



Gambar 2. Alur Pemrosesan Data

Pemrosesan data merupakan metode yang digunakan untuk menyiapkan set data sebelum dilakukannya pemodelan. Tahapan pemrosesan data termasuk salah satu tahapan yang penting dilakukan khususnya dalam pemodelan data tekstual karena bertujuan untuk mengekstrak data ke dalam format yang sama atau serupa. [9] Tahapan pemrosesan data yang diimplementasikan meliputi:

a. *Cleansing*, pembersihan data tekstual dengan cara menghilangkan sekumpulan karakter atau symbol khusus selain huruf dari data.
b. *Case folding*, melibatkan pengubahan data tekstual ke dalam bentuk karakter huruf kecil.
c. *Normalization*, proses normalisasi data yang memiliki format salah karena salah ketik atau *typo* maupun kata tidak baku ke dalam format standar.
d. *Stemming,* proses mengubah kata pada data ke dalam bentuk kata dasar, seperti “menolak” menjadi “tolak”
e. *Penggabungan negasi,* menggabungkan dua kata yang berawalan kata ‘tidak’, ‘belum’, ‘jangan’, ‘kurang’ yang biasanya dikonotasikan atau bermakna negatif.
f. *Stopword removal*, penghapusan kata-kata dalam data yang memiliki makna minim, seperti kata ganti orang, kata penghubung, dan sebagainya.

2.3. *Term Frequency-Inverse Document Frequency* (TF-IDF)

TF-IDF merupakan teknik ekstraksi fitur kata ke dalam bobot angka yang mencerminkan seberapa besar kepentingan kata tersebut dalam suatu dokumen relatif terhadap koleksi dokumen lainnya. Secara sederhana perhitungan pembobotan TF-IDF dapat dianalogikan: "Jika suatu kata sering atau terus muncul dalam dokumen pertama, namun kata tersebut tidak muncul pada dokumen kedua atau dokumen lainnya, maka kata tersebut diartikan memiliki makna yang penting untuk dokumen pertama." Teknik ekstraksi fitur ini dilakukan dengan menggabungkan dua metrik, yakni *Term Frequency* (TF) dan *Inverse Document Frequency* (IDF).

Secara matematis, TF-IDF dapat direpresentasikan sebagai berikut. [10]

$$w_{i,j} = tf_{i,j} \times \log(\frac{N}{df_i}) \quad (1)$$

- $tf_{i,j}$ merupakan TF (*Term Frequency*), frekuensi kemunculan kata i dalam dokumen j. Dihitung dengan membagi jumlah kemunculan kata i dalam dokumen j dengan total seluruh kata dalam dokumen j.
- df_i merupakan frekuensi atau jumlah dokumen yang mengandung kata i yang berperan mengukur pentingnya istilah di setiap korpus.
- N merupakan jumlah dokumen.

2.4. *Support Vector Machine* (SVM)

Support Vector Machine (SVM) merupakan algoritma pengenalan pola yang dapat digunakan dalam menyelesaikan masalah klasifikasi, regresi, dan deteksi linear. Secara sederhana konsep SVM merupakan upaya untuk menemukan *hyperplane* terbaik yang berperan sebagai pemisah antara dua kelas atau lebih pada suatu ruang input. *Hyperplane* terbaik ditemukan dengan cara mengukur margin atau jarak antara *hyperplane* dengan pola kelas terdekat. [11] Dalam permasalahan klasifikasi biner atau dua kelas, *hyperplane* atau pemisah dua kelas dapat diilustrasikan secara garis linear. Namun pada kenyataannya, sebaran data cenderung beragam sehingga cukup sulit jika dipisahkan secara linear. Untuk mengatasi hal ini, SVM memperkenalkan fungsi kernel, yang memungkinkan transformasi ruang data asli menjadi ruang data baru dengan dimensi yang lebih tinggi dan lebih mudah dipisahkan berdasarkan kategori data yang telah ditentukan. Terdapat empat jenis kernel yang diperkenalkan pada SVM, diantaranya kernel *linear*, *radial basis function* (RBF), *sigmoid*, dan *polynomia*l. Dimana setiap kernel memiliki parameter yang perlu dioptimasi untuk mendapatkan kinerja model terbaik. [12]

Tabel 1. Kernel Suppor Vector Machine (SVM)

No	Fungsi Kernel	Formula	Parameter
1	Linear	$K(X_i, X_j) = (X_i, X_j)$	C dan γ
2	RBF	$K(X_i, X_j) = \exp(-\gamma \left\|\|X_i, X_j\right\|\|^2 + C)$	C dan γ
3	Sigmoid	$K(X_i, X_j) = \tanh(\gamma (X_i, X_j + r)$	C, r, dan γ
4	Polynomial	$K(X_i, X_j) = (\gamma(X_i, X_j) + r)^d$	C, γ, r, dan d

Dimana,

- X_i merupakan *support vector data*, dengan $i = 1, 2,3,4,5, \ldots..$
- X_j merupakan nilai data ke-j
- C merupakan *cost*
- γ merupakan *gamma*
- r merupakan *coefficient* atau koefesien
- d merupakan *degree*

p-ISSN: 2301-5373
e-ISSN: 2654-5101

2.5. *Hyperparameter Tuning*

Dalam konteks pembelajaran mesin, *hyperparameter tuning* merupakan proses menetapkan parameter model sebelum memulai proses pembelajaran. Parameter model mengacu pada bobot dan koefisien yang diturunkan dari data oleh algoritma. Setiap algoritma memiliki set *hyperparameter* yang berbeda. Proses *hyperparameter tuning* berperan dalam memaksimalkan kinerja model pada data validasi.

2.6. *K-Fold Cross Validation*

Cross validation atau validasi silang merupakan salah satu metode yang digunakan dalam analisis data untuk memvalidasi atau mengestimasi kinerja model pembelajaran statistik. Dengan kata lain, *cross validation* berperan untuk mengevaluasi seberapa baik mesin statistik menggeneralisasi data baru yang tidak digunakan dalam pelatihan model. Dalam *k-fold cross validation*, set data dibagi secara acak berdasarkan k kelipatan atau kelompok dengan ukuran yang kurang lebih sama. Salah satu subset digunakan sebagai data pengujian atau data validasi dan subset lainnya digunakan sebagai data pelatihan. Metode ini sangat akurat digunakan dalam mengevaluasi kinerja model, tetapi memerlukan lebih banyak sumber daya komputasi. Dalam praktiknya, pilihan jumlah k dapat disesuaikan tergantung pada ukuran set data, meskipun jumlah k=5 dan k=10 merupakan pilihan kelipatan yang paling umum digunakan. [13]

2.7. *Confusion Matrix*

Matriks digunakan untuk menilai performa model *machine learning*. Dalam konteks klasifikasi, hasil perhitungan performa model dapat dirangkum dalam *confusion matrix*, yang membagi hasil tes data sampel ke dalam empat kategori tergantung label data sebenarnya (*true label*) *d*an label data hasil prediksi (*predicted tabel*). [14]

Tabel 2. *Confusion matrix*

		True Label	
		Positive	**Negative**
Predicted Label	**Positive**	**TP**	**FP**
	Negative	**FN**	**TN**

Dalam evaluasi model klasifikasi, interpretasi hasil digambarkan dalam *confusion matrix* yang terdiri dari *True Positive* (TP), *True Negative* (TN), *False Positive* (FP), dan *False Negative* (FN). TP merujuk pada jumlah data sentimen positif yang benar-benar diprediksi positif oleh model. TN mengindikasikan jumlah data sentimen negatif dan diprediksi negatif oleh model. FP mengacu pada jumlah data sentimen yang sebenarnya negatif tetapi salah diprediksi positif oleh model, sedangkan FN menunjukkan jumlah data sentimen yang sebenarnya positif tetapi salah diprediksi sebagai negatif oleh model. Dalam konteks interpretasi, tujuan utamanya adalah mengoptimalkan jumlah TP dan TN serta mengurangi jumlah FP dan FN sehingga diperoleh model *machine learning* dengan performa yang optimal. [14]

Matriks *precision, recall, F1-score*, dan *accuracy* akan digunakan dalam penelitian ini untuk menilai performa pada model SVM, yang direpresentasikan sebagai berikut.

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

$$F_1 - score = \frac{2TP}{2TP+FP+FN} \quad (4)$$

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (5)$$

2.8. *Latent Dirichlet Allocation* (LDA)

LDA merupakan salah satu metode pemodelan topik yang populer. LDA dapat digunakan untuk menguraikan makna simantik teks untuk menentukan topik utama. Konsep dasar LDA adalah setiap topik diwakili oleh sebuah kata yang mirip dengan kata lainnya dan setiap dokumen dapat diwakili oleh sejumlah topik. Setiap dokumen akan direpresentasikan ke dalam sebuah topik yang berisi istilah-istilah dalam dokumen tersebut menurut LDA. Tidak penting untuk mengubah kata-kata dokumen atau perubahan konten karena LDA memandang setiap dokumen sebagai sekumpulan kata. [15]

Secara matematis, distribusi dari variable laten (topik dan penugasan topik) dan variable yang diamati (kata-kata) dalam satu dokumen dapat dapat direpresentasikan sebagai berikut. [15]

$$\mathrm{p}(\theta, \mathrm{z}, \mathrm{w} \mid \alpha, \beta) = \mathrm{p}(\theta \mid \alpha)(\prod_{n=1}^{\mathrm{N}} \mathrm{p}\,(\mathrm{z}_n \mid \theta)\mathrm{p}(\mathrm{w}_n \mid \mathrm{z}_n, \beta) \quad (6)$$

Kemudian distrubusi marginal dari satu dokumen diperoleh dengan mengintegrasikan θ dan menjumlahkan z:

$$p(w|\alpha,\beta) = \int p(\theta|\alpha)\ (\prod_{n=1}^{N} \sum_{z_n} p(z_n|\theta)p(w_n|z_n,\beta)) \quad (7)$$

Terakhir dengan dengan mengalikan probabilitas distribusi marginal diperoleh probabilitas dari satu corpus:

$$p(D|\alpha,\beta) = \prod_{d=1}^{M} \int p(\theta_d|\alpha)\,(\prod_{n=1}^{N_d} \sum_{z_{dn}} p(z_{dn}|\theta)p(w_{dn}|z_{dn},\beta)) \quad (8)$$

Dimana:

- θ adalah distribusi topik dokumen.
- z adalah penugasan untuk setiap kata.
- w adalah kata yang diamati.
- α adalah parameter distribusi topik per dokumen.
- β adalah parameter distribusi kata per topik.
- D adalah panjang dokumen.
- θ_d adalah variable tingkat dokumen, disampel satu kali per dokumen.
- z_{dn} dan w_{dn} adalah variable tingkat kata, disampe satu kali untuk setiap kata dalam setiap dokumen.

2.9. *Topic Coherence*

Topic coherence merupakan salah satu pengukuran pada model pemodelan topik yang membantu dalam membedakan topik yang ditafsirkan secara simantik dari topik yang merupakan hasil dari inferensi statistik. *Topic coherence* didefinisikan sebagai rata-rata dari *coherence score* tertentu untuk setiap pasangan kata. Misalnya, untuk topik dengan kata – kata [politik, negera, Indonesia], akan dihitung nilai koherensi untuk setiap pasangan kata (politik, negara), (politik, Indonesia), dan (negara, Indonesia), kemudian hasil nilai koherensi dihitung berdasarkan nilai rata – rata setiap pasang kata tersebut untuk menadapatkan nilai koherensi topik. Secara matematis dapat direpresentasikan sebagai berikut. [16]

$$TopicCoherence(z, D) = mean\{score(w_i, w_j, \in)\} \quad (9)$$

Dimana:

- z adalah sebuah topik (sekumpulan kata yang mendeskripsikan z)
- D adalah koleksi dokumen (sekumpulan dokumen)
- $score$ adalah ukuran koherensi antara sepasang kata
- V merepresentasikan seluruh kosakata yang ada di D
- w_i, w_j merepresentasikan sepasang kata yang mendeskripsikan topik
- $Term\ epsilon$, digunakan sebagai smoothing value tergantung pada sifat dataset dan mencegah terjadinya nilai ekstrim.

3. Hasil dan Pembahasan

3.1. Hasil Analisis Sentimen dengan SVM

Pemodelan analisis sentimen dengan SVM memperhatikan beberapa *hyperparameter tuning*, diantaanya fungsi kernel, C, *gamma*, dan d*egree* yang bertujuan untuk menentukan parameter terbaik dari model SVM. Fungsi kernel yang digunakan dalam skenario pengujian ini adalah *linear*, RBF, *sigmoid*, dan *polynomial*. Untuk *hyperparameter* "C" akan digunakan rentang 0,1; 1,0; dan 10,0. Untuk *hyperparameter* γ atau *gamma* adalah 0,0001; 0,001; 0,01; 1,0; 0,1; dan 10,0. Serta untuk untuk *hyperparameter* $degree\ (d)$ adalah 1; 2; 3; 4; dan 5.

Dataset akan dibagi menjadi dua, yaitu data *training* dan testing, dengan jumlah 2000 data *training* dan 492 data *testing.* Proses *hyperparameter tuning* menggunakan teknik *grid search* untuk memungkinkan eksplorasi berbagai *hyperparameter*. Pengujian model awal dilakukan pada data *training* menggunakan 10-*fold cross validation*, artinya data dibagi menjadi 10 subset (bagian). Kombinasi 9 subset yang berbeda digabungkan dan digunakan sebagai data *training* dan 1 subset sisa digunakan sebagai data *validation*. Dari masing - masing kernel nanti akan dipilih hasil *hyperparameter tuning* terbaik, dengan memperhatikan nilai rata - rata akurasi *train* dan *validation*, *precision*, *recall*, dan *f-1 score*. Setiap kernel dengan konfigurasi *hyperparameter* yang optimal telah dipilih berdasarkan kinerja model pada data validasi seperti pada Tabel 3.

Tabel 3. Perbandingan Hasil Setiap Kernel pada Proses Tuning

Kernel	C	γ	d	Accuracy	Precision	Recall	F-1 Score
linear	1,0	-	-	0,849	0,8497	0,849	0,8489
rbf	10,0	1,0	-	0,8615	0,862	0,8615	0,8615
sigmoid	1,0	1,0	-	0,844	0,8448	0,844	0,8439
polynomial	1,0	-	2	0,8625	0,8632	0,8625	0,8624

Proses pengujian dilanjutkan pada data *testing* bertujuan untuk mengevaluasi kinerja model yang telah di-*tuning* dengan menggunakan data yang belum pernah dilihat sebelumnya, sehingga dapat memberikan gambaran terhadap kemampuan model dalam memprediksi data baru. Untuk memilih model SVM dengan kernel yang paling sesuai dengan analisis sentimen pada kendaraan listrik mengacu pada model dengan akurasi*, precision*, *recall*, dan *f-1 score testing* tertinggi. Adapun hasil pengujian model setiap jenis kernel pada data *testing* ditunjukkan pada Tabel 4. Dalam pengujian ini, hasil menunjukkan model SVM dengan kernel *polynomial* dengan C=1.0 dan derajat=2 memberikan hasil terbaik dan konsisten dengan akurasi sebesar 0,8618; presisi 0,86; recall 0,86; dan F-1 score 0,86. Meskipun demikian, model SVM kernel RBF dengan C=10.0 dan gamma=1.0 juga memberikan performa yang baik dan konsisten pada akurasi, presisi, recall, dan F-1 score dengan nilai sebesar 0,85. Namun, performa kernel RBF pada proses *tuning* dengan pengujian data *testing* mengalami sedikit penurunan. Sedangkan untuk model SVM kernel *linear* dan sigmoid menunjukkan performa yang hampir serupa dengan akurasi, presisi, recall, dan F-1 score sekitar 0,84.

Tabel 4. Hasil Pengujian Data Testing pada Setiap Kernel

Kernel	C	γ	d	Accuracy	Precision	Recall	F-1 Score
linear	1,0	-	-	0,841	0,84	0,84	0,84
rbf	10,0	1,0	-	0,8536	0,85	0,85	0,85
sigmoid	1,0	1,0	-	0,8435	0,84	0,84	0,84
polynomial	**1,0**	**-**	**2**	**0,8618**	**0,86**	**0,86**	**0,86**

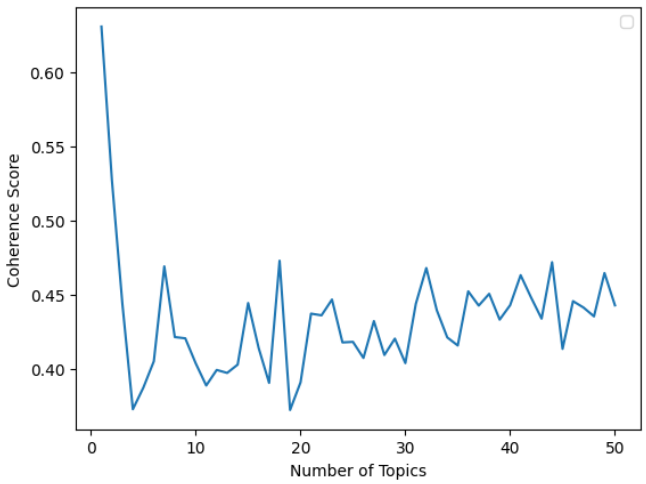
Adapun hasil analisis sentimen pada data *testing* dapat dilihat pada *confusion matrix* hasil prediksi pada Tabel 5. Dari 492 data uji, sudah terdapat cukup banyak prediksi yang dihasilkan akurat sesuai dengan data aktual baik pada data dengan sentiment positif maupun negatif, yakni jumlah TP sebesar 211 dan jumlah TN sebesar 213. Hal ini menunjukkan model SVM yang telah dioptimasi berdasarkan tuning parameter baik dan andal dalam memprediksi sentimen terkait kendaraan listrik.

Tabel 5. *Confusion Matrix* Hasil Prediksi Analisis Sentimen Kendaraan Listrik

		True Label	
		Positive	Negative
Predicted Label	Positive	211	33
	Negative	35	213

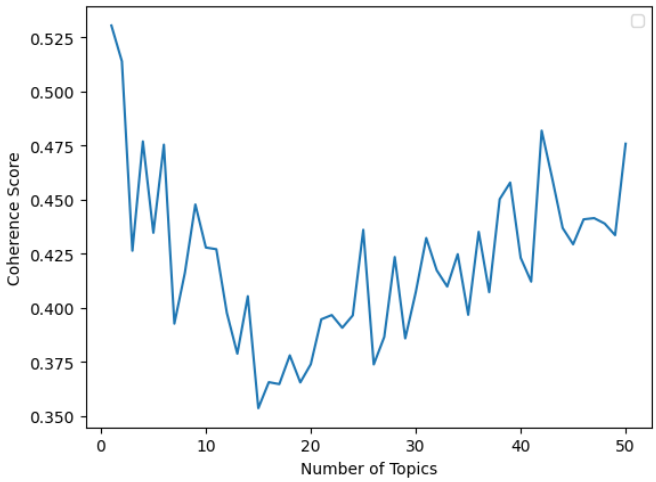
3.2. Hasil Pemodelan Topik dengan LDA

Dalam pemodelan topik dengan LDA diperlukan adanya uji koherensi untuk memperoleh model terbaik. Dalam menentukan memperoleh nilai koherensi yang maksimal terdapat beberapa parameter yang dapat di-*tuning* salah satunya adalah jumlah atau banyak topik. Representasi banyak topik pada model LDA yang digunakan dalam rentang topik 1 hingga 50. Nantinya akan dipilih banyak topik yang paling baik dengan mempertimbangkan nilai koherensinya.



Gambar 3. Grafik Sebaran Nilai Koherensi pada Sentimen Positif

p-ISSN: 2301-5373
e-ISSN: 2654-5101



Gambar 4. Grafik Sebaran Nilai Koherensi pada Sentimen Negatif

Ditujukan pada Gambar 3, grafik nilai koherensi pada pemodelan topik sentimen positif dengan menunjukkan nilai yang tinggi pada jumlah topik satu dan adanya peningkatan nilai koherensi seiring dengan meningkatnya jumlah topik (*flatting out*). Jika nilai koherensi pada satu topik sangat besar, itu dapat menunjukkan bahwa model hanya mencoba menemukan pola umum dari seluruh korpus, bukan topik – topik spesifik secara terperinci. Sedangkan ketika meningkatnya jumlah topik, memungkinkan model untuk memiliki lebih banyak fleksibilitas dalam menemukan pola – pola dalam data. Namun, peningkatan jumlah topik sering kali menimbulkan tumpang tindih antar topik. Sehingga jumlah topik yang memiliki nilai koherensi tertinggi sebelum terjadi *flatting out*, yaitu jumlah topik sebanyak 3 topik dengan nilai koherensi 0,4437 dipilih sebagai model LDA paling stabil dan baik untuk sentimen positif. Sedangkan pada Gambar 4, grafik nilai koherensi pada pemodelan topik sentimen negatif dengan 150 iterasi juga menunjukkan nilai yang tinggi pada jumlah topik satu dan adanya *flatting out*. Sehingga jumlah topik yang memiliki nilai koherensi tertinggi sebelum terjadi *flatting out*, yaitu jumlah topik sebanyak 4 topik dengan nilai koherensi 0,47698 dipilih sebagai model LDA pada sentimen negatif.

Adapun hasil pemodelan topik pada data kendaraan listrik dengan sentiment positif dan negatif ditunjukkan pada Tabel 6 dan 7.

Tabel 6. Hasil Pemodelan Topik Sentimen Positif

Topik	Kata
1	'0.006*"dukung" + 0.005*"mantap" + 0.005*"alih" + 0.005*"terus" + ' '0.005*"baik" + 0.004*"solusi" + 0.004*"polusi" + 0.004*"suka" + ' '0.004*"lanjut" + 0.003*"suara"'
2	'0.010*"tuju" + 0.007*"lebih" + 0.007*"bagus" + 0.006*"orang" + 0.006*"beli" ' '+ 0.005*"banyak" + 0.005*"polusi" + 0.005*"kurang" + 0.005*"buat" + ' '0.005*"minyak"'
3	'0.010*"minyak" + 0.010*"bahan" + 0.009*"bakar" + 0.008*"indonesia" + ' '0.007*"pakai" + 0.006*"negara" + 0.006*"jadi" + 0.006*"lebih" + ' '0.006*"banget" + 0.005*"kalau"'

Pada data dengan sentimen positif, Topik 1 dapat diinterpretasikan bahwa dibahas tentang dukungan masyarakat akan transisi kendaraan listrik. Kata yang menekankan dukungan diwakilkan dengan kata “dukung”, “mantap”, “suka”, “terus”, “baik”, dan “lanjut”.

Topik 2, dapat diinterpretasikan tentang minat masyarakat terhadap kendaraan listrik yang diwakilkan dengan kata “bagus”, “orang”, “banyak”, dan “beli” menekankan tentang persepsi positif, ketertarikan, dan tindakan konsumtif masyarakat untuk membeli kendaraan listrik sebagai alternatif.

Topik 3, dapat diinterpretasikan tentang peran Indonesia dalam transisi ke kendaraan listrik. Diwakilkan dengan kata “Indonesia”, “pakai”, “negara”, “jadi”, dan “kalau” yang menggambarkan pentingnya Indonesia menjadi salah satu negara pelopor kendaraan listrik. Selain itu, kata “minyak”, “bahan”, “bakar” menyoroti kebutuhan untuk mengurangi ketergantungan pada bahan bakar fosil.

Tabel 7. Hasil Pemodelan Topik Sentimen Negatif

Topik	Kata
1	'0.012*"orang" + 0.010*"kaya" + 0.009*"beli" + 0.008*"buat" + 0.007*"bahan" ' '+ 0.007*"bakar" + 0.006*"lebih" + 0.006*"baterai" + 0.006*"pribadi" + ' '0.006*"mampu"'
2	'0.006*"untung" + 0.006*"orang" + 0.005*"usaha" + 0.005*"rakyat" + ' '0.005*"batu bara" + 0.005*"bangkit" + 0.005*"sumber" + 0.004*"bodoh" + ' '0.004*"listrik" + 0.004*"beli"'
3	'0.005*"buat" + 0.005*"juta" + 0.004*"beli" + 0.004*"rakyat" + 0.004*"orang" ' '+ 0.004*"lebih" + 0.004*"jadi" + 0.004*"kalau" + 0.004*"baik" + ' '0.004*"subsidi"'
4	'0.008*"mahal" + 0.006*"harga" + 0.006*"rakyat" + 0.005*"orang" + ' '0.005*"kalau" + 0.005*"beli" + 0.004*"lebih" + 0.004*"buat" + 0.004*"butuh" ' '+ 0.004*"bukan"'

Pada data dengan sentimen negatif, Topik 1, dapat diinterpretasikan tentang sentimen masyarakat bahwa subsidi kendaraan listrik tidak tepat sasaran. Tidak tepat sasaran tersebut diwakilkan dengan kata “orang”, “kaya”, “mampu”, “beli”, “buat”, “lebih” yang menyoroti persepsi bahwa subsidi kendaraan listrik hanya diperuntukkan atau hanya dapat dinikmati oleh orang kaya atau orang yang mampu.

Topik 2, dapat diinterpretasikan sebagai topik yang membahas subsidi kendaraan listrik hanyak untuk kepentingan atau menguntungkan pengusaha dan tantangan infrastruktur.

Topik 3, diinterpretasikan mengenai kendala finansial dalam pembelian kendaraan listrik. Kata “buat”, “juta”, “beli”, “rakyat”, dan “orang” menggambarkan tantangan finansial yang dihadapi masyarakat dalam membeli kendaraan listrik yang harganya sangat tinggi, bahkan mencapai jutaan rupiah.

Topik 4 memiliki interpretasi yang hampir mirip dengan topik 3 membahas mengenai persepsi masyarakat yang melihat kendaraan listrik sebagai produk yang mahal. Kata “mahal” dan “harga” menyoroti biaya tinggi yang harus dikeluarkan untuk membeli kendaraan listrik.

4. Kesimpulan

Penerapan metode *machine learning* dalam analisis sentimen dan pemodelan topik dapat digunakan dengan cukup baik pada data opini kendaraan listrik. Dari 492 data uji opini masyarakat tentang kendaraan listrik yang diperoleh melalui media sosial Twitter dan YouTube sebanyak 244 data berlabel positif dan 248 data berlabel negatif. Dari proses analisis sentimen dengan model SVM menghasilkan performa terbaik dan konsisten dengan parameter parameter kernel polynomial dengan nilai C=1.0 dan degree=2, mencapai nilai akurasi, presisi, recall, dan F1-score masing – masing sebesar 0,86 atau dalam persentase sebesar 86%.

Pada pemodelan topik data kendaraan listrik menggunakan model LDA. Pengelompokkan topik dilakukan untuk masing – masing jenis sentimen positif dan negatif, dengan uji coba parameter jumlah topik. Model LDA terbaik untuk sentimen positif adalah model dengan jumlah topik 3 dan menghasilkan nilai coherence sebesar 0,4437. Data dengan sentimen positif banyak membahas tentang dukungan dan ketertarikan masyarakat tentang kendaraan listrik. Sedangkan model LDA terbaik untuk sentimen negatif adalah model dengan jumlah topik 4 dan menghasilkan nilai coherence sebesar 0,47698. Data dengan sentimen negatif banyak membahas tentang subsidi kendaraan listrik yang dinilai kurang tepat sasaran, infrastruktur yang belum memadai, dan persepsi masyarakat tentang kendaraan listrik sebagai produk yang mahal.

p-ISSN: 2301-5373
e-ISSN: 2654-5101

Referensi

[1] R. A. Subekti, H. Sudibyo, V. Susanti, H. M. Saputra, and A. Hartanto, *Peluang dan Tantangan Pengembangan Mobil Listrik Nasiona*, Pertama. Jakarta: LIPI Press, 2014.
[2] Nissan, "Kisah Sejarah Mobil Listrik, Dari Ide Menjadi Mobil Masa Depan," *nissan.co.id*, 2022. https://nissan.co.id/new-press/artikel/kisah-sejarah-mobil-listrik-dari-ide-menjadi-mobil-masa-depan/ (accessed Mar. 21, 2023).
[3] S. Sandya, "Banyak Masyarakat Belum Minat Pakai Mobil Listrik, Ini Alasannya," *dataindonesia.id*, 2022. https://dataindonesia.id/sektor-riil/detail/banyak-masyarakat-belum-minat-pakai-mobil-listrik-ini-alasannya (accessed Mar. 25, 2023).
[4] Yusuf, "Beli Motor Listrik Dapat Bantuan Pemerintah Rp7 Juta, Ini Syaratnya!," *kominfo.go.id*, 2023. https://www.kominfo.go.id/content/detail/48097/beli-motor-listrik-dapat-bantuan-pemerintah-rp7-juta-ini-syaratnya/0/berita (accessed Mar. 25, 2023).
[5] P. Indonesia, *Peraturan Menteri Keuangan Nomor 38 Tahun 2023*. Jakarta: Menteri Keuangan Republik Indonesia, 2023.
[6] A. Erfina and R. A. Lestari, "Sentiment Analysis of Electric Vehicles using the Naïve Bayes Algorithm," *Sistemasi*, vol. 12, no. 1, p. 178, 2023, doi: 10.32520/stmsi.v12i1.2417.
[7] M. R. Fahlevvi, "Sentiment Analysis And Topic Modeling on User Reviews of Online Tutoring Applications Using Support Vector Machine and Latent Dirichlet Allocation," *Knowbase Int. J. Knowl. Database*, vol. 2, no. 2, p. 142, 2022, doi: 10.30983/knowbase.v2i2.5906.
[8] A. Santoso, A. Nugroho, and A. S. Sunge, "Analisis Sentimen Tentang Mobil Listrik Dengan Metode Support Vector Machine Dan Feature Selection Particle Swarm Optimization," *J. Pract. Comput. Sci.*, vol. 2, no. 1, pp. 24–31, 2022, doi: 10.37366/jpcs.v2i1.1084.
[9] S. Celik and S. Gulsecen, "Data Pre-processing in Text Mining CHAPTER 7 DATA PRE-PROCESSING IN TEXT MINING," no. December, pp. 122–144, 2020, doi: 10.26650/B/ET06.2020.011.07.
[10] S. Vajjala *et al.*, *Practical Natural Language Processing*. O'Reilly Media, Inc., 2020.
[11] A. . Nugroho, A. . Witarto, and D. Handoko, *Support Vector Machine – Teori dan Aplikasinya dalam Bioinformatika*. 2003.
[12] M. A. Nanda, K. B. Seminar, D. Nandika, and A. Maddu, "A Comparison Study of Kernel Functions in the Support Vector Machine and A Comparison Study of Kernel Functions in the Support Vector Machine and Its Application for Termite Detection," no. January, 2018, doi: 10.3390/info9010005.
[13] O. A. Montesinos López, M. L. A, and C. J. Cham, *Multivariate Statistical Machine Learning Methods for Genomic Prediction*. 2022.
[14] G. Varoquaux and O. Colliot, *Evaluating machine learning models and their diagnostic value*. O. Colliot (Ed.), Machine Learning for Brain Disorders, Springer, 2022.
[15] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *J. Mach. Learn. Res. 3*, pp. 933–1022, 2003, doi: 10.1016/B978-0-12-411519-4.00006-9.
[16] A. R. Pasquali, "Automatic Coherence Evaluation Applied to Topic Models," p. 82, 2016.

This page is intentionally left blank.