PENINGKATAN MODEL KLASIFIKASI SENTIMEN PUBLIK DI YOUTUBE CNBC INDONESIA TERHADAP MOBIL LISTRIK MENGGUNAKAN ALGORITMA SUPPORT VEKTOR MACHINE

Sehabudin [1], Ade Irma Purnamasari [2], Agus Bahtiar [3], Edi Wahyudin [4]
[1,2] Jurusan Teknik Informatika, STMIK IKMI Cirebon
[3] Jurusan Sistem Informasi, STMIK IKMI Cirebon
[4] Jurusan Komputerisasi Akuntasnsi, STMIK IKMI Cirebon
Jl. Perjuangan No.10B, Karyamulya, Kec. Kesambi, Kota Cirebon, Jawa Barat 45135
sehabb202@gmail.com

ABSTRAK

Pertumbuhan teknologi kendaraan listrik di Indonesia memicu beragam opini publik, khususnya di platform media sosial seperti YouTube. CNBC Indonesia, sebagai salah satu sumber informasi terpercaya, sering menjadi pusat diskusi terkait isu ini. Penelitian ini bertujuan untuk meningkatkan model klasifikasi sentimen publik terhadap kendaraan listrik menggunakan algoritma Support Vector Machine (SVM), yang dikenal memiliki performa unggul dalam klasifikasi berbasis teks. Data penelitian berupa komentar publik dari kanal YouTube CNBC Indonesia yang dikumpulkan dari video-video bertema kendaraan listrik selama periode tertentu. Tahapan penelitian meliputi preprocessing data seperti tokenisasi, penghapusan stopwords, stemming, pembobotan dengan Term Frequency-Inverse Document Frequency (TF-IDF), dan implementasi algoritma SVM untuk klasifikasi sentimen menjadi positif dan negatif. Evaluasi model dilakukan menggunakan metrik akurasi, presisi, recall, dan F1-score, serta membandingkan hasilnya dengan algoritma lain seperti Naïve Bayes. Hasil penelitian menunjukkan bahwa algoritma SVM memiliki performa terbaik dengan akurasi mencapai 93% dan F1-score yang konsisten tinggi. Sebagian besar sentimen publik terhadap kendaraan listrik bersifat negatif, meskipun terdapat kritik terkait infrastruktur dan biaya yang masih menjadi tantangan. Keberhasilan algoritma SVM menunjukkan potensinya untuk analisis teks yang lebih kompleks di masa depan, meskipun penelitian ini menghadapi tantangan seperti bias data dan perlunya memperluas dataset.

***Kata kunci**:Sentimen Publik, Mobil Listrik, YouTube, Support Vector Machine, Klasifikasi*

1. PENDAHULUAN

Perkembangan pesat di bidang informatika telah memberikan dampak signifikan pada berbagai sektor kehidupan, termasuk teknologi, bisnis, dan pendidikan. Inovasi teknologi yang semakin canggih, seperti artificial intelegen (AI), Machine Learning (ML), dan Natural Languange Processing (NLP), memungkinkan penerapan teknik komputasi dalam menyelesaikan berbagai permasalahan yang kompleks. Salah satu tren terkini adalah implementasi teknologi tersebut dalam analisis sentimen, khususnya dalam konteks media sosial yang memuat opini masyarakat terhadap isu-isu tertentu. Dalam hal ini, topik yang menjadi perhatian publik saat ini adalah mobil listrik, yang semakin banyak diperbincangkan di berbagai platform, termasuk YouTube. Konten video yang diproduksi oleh CNBC Indonesia, misalnya, sering kali menghadirkan diskusi seputar mobil listrik, yang ditanggapi oleh masyarakat melalui komentar-komentar yang mencerminkan sentimen mereka.

Namun, berbagai masalah yang cukup kompleks muncul saat menganalisis sentimen komentar media sosial. Keberanekaragaman bahasa yang digunakan oleh masyarakat merupakan masalah utama. Ini terutama berlaku untuk bahasa informal, yang seringkali sulit dipahami oleh model pemrosesan bahasa alami tradisional. Dengan variasi emosi dan sentimen yang diungkapkan dalam komentar, mulai dari sentimen positif, dan negatif, tantangan ini menjadi lebih rumit. Selain itu, ironi atau sarkasme sering digunakan, yang dapat mengaburkan sentimen. Penting bagi penelitian ini betapa tidak akuratnya model tradisional dalam mengidentifikasi pola sentimen yang kompleks. Untuk mendapatkan hasil yang relevan dan dapat diandalkan dalam analisis opini masyarakat tentang mobil listrik, sangat penting untuk menginterpretasikan perasaan publik dengan benar. masalah utama yang sering kali muncul adalah akurasi model klasifikasi sentimen yang masih rendah dalam mengenali pola opini yang bermacam-macam. Meskipun algoritma pembelajaran mesin telah digunakan secara luas dalam klasifikasi sentimen, terdapat kesenjangan antara kebutuhan untuk mencapai akurasi tinggi dan keterbatasan yang ada pada model-model tradisional yang tidak mampu menangkap kompleksitas data yang ada. Oleh karena itu, menemukan metode yang lebih cocok dan tepat sangat penting untuk mengatasi masalah dalam analisis sentimen mobil listrik di platform seperti YouTube.

Tujuan utama penelitian ini adalah untuk membuat dan menguji model klasifikasi sentimen dengan menggunakan algoritma Support Vector Machine (SVM) terhadap komentar masyarakat tentang mobil listrik di YouTube.Adapun tujuan lainnya untuk membuat model yang akurat dan efisien untuk mengidentifikasi sentimen positif, negatif yang berkaitan dengan persepsi masyarakat terhadap

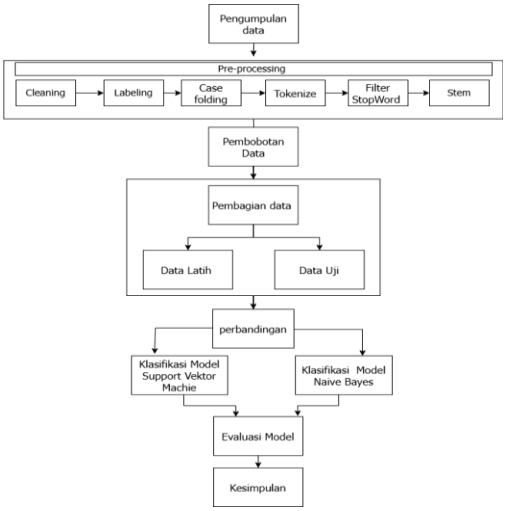
inovasi kendaraan ramah lingkungan ini. Penelitian ini penting karena dapat mengisi celah pengetahuan tentang analisis sentimen di media sosial, khususnya tentang transportasi berkelanjutan di Indonesia. Dalam penelitian ini diharapkan dapat memberi pembuat kebijakan dan industri otomotif wawasan yang lebih mendalam tentang persepsi masyarakat terhadap mobil listrik dan faktor-faktor yang memengaruhi persepsi tersebut. Penelitian ini juga dapat membantu mengembangkan metode analisis sentimen di bidang informatika. Hasilnya dapat digunakan sebagai dasar untuk penelitian lebih lanjut di bidang yang serupa.

2. TINJAUAN PUSTAKA

Pada tahap permulaan penelitian ini, literatur review dilaksanakan untuk mencari penelitian yang terkait yang dapat mendukung penelitian ini. Kajian ulang literatur menunjukkan studi analisis sentimen, seperti penelitian [1]yang menggunakan algoritma SVM untuk klasifikasi penyakit cacar monyet di negara non-endemik. Pendekatan KDD diterapkan dengan empat kernel: linear, RBF, sigmoid, dan polynomial. Hasilnya, kernel polynomial memberikan akurasi terbaik sebesar 75%, diikuti linear 70,5%, RBF 66%, dan sigmoid 45%. Algoritma SVM dipilih karena fleksibilitasnya dalam menangkap pola non-linear melalui ruang fitur yang tinggi. Selain itu, nilai ROC AUC kernel polynomial sebesar 0,81 mengindikasikan performa model yang baik, sehingga relevan untuk dikembangkan lebih lanjut.

Penelitian lebih lanjut [2]) menganalisis sentimen di Twitter terkait kebijakan kenaikan harga BBM pada September 2022 menggunakan algoritma SVM dengan kernel linear, RBF, dan polynomial. Hasilnya menunjukkan reaksi negatif dominan, dengan kernel RBF mencatat kinerja terbaik sebesar 87% menggunakan TF-IDF, lebih akurat dibandingkan BoW. Teknik SMOTE meningkatkan kinerja kernel polynomial sebesar 2% pada data split 70:30 dan 80:20, memberikan wawasan untuk kebijakan lebih responsif.

3. METODE PENELITIAN



Gambar 1. Tahapan Metode Penlitian

Penelitian ini dilakukan dengan langkah-langkah sistematis melaui pendekatan kuantitatif dengan eksperimen. Tujuan dari penelitian ini adalah untuk membangun dan meningkatkan model klasifikasi sentimen publik di youtube CNBC Indonesia menggunakan algoritma Support Vector Machine (SVM).

3.1. Pengambilan Data

Pada tahap ini, data sentimen publik yang relevan dikumpulkan dari kolom komentar YouTube CNBC Indonesia terkait topik mobil listrik. Data diambil dari video-video yang membahas mobil listrik yang diposting dalam periodeter tentu. Pengpumpulan data meggunakan Teknik carwling data menggunakan aplikasi python dengan memasukan APIkey youtube.

3.2. Pre-processing

Data komentar yang diambil dari YouTube perlu melalui proses praproses agar dapat dianalisis dengan lebih baik. Tahap ini melibatkan beberapa langkah:

a. Cleaning (Pembersihan Data): Menghapus karakter atau kata yang tidak relevan seperti tanda baca, URL, dan angka.
b. Labeling : Memberikan label pada data seperti positif dan negatif
c. Case Folding: Mengubah seluruh teks menjadi huruf kecil untuk konsistensi.
d. Tokenization: Memecah kalimat menjadi kata-kata yang terpisah untuk memudahkan analisis.
e. Stopword Removal: Menghapus kata-kata umum yang tidak memiliki makna signifikan (seperti "dan", "di", "ke").
f. Stemming: Mengubah kata-kata ke bentuk dasarnya untuk menyederhanakan analisis.

3.3. Pembobotan data

Setelah data selesai diproses, tahap ini mengubah data ke dalam bentuk yang dapat dianalisis oleh algoritma. Dalam penelitian ini, teknik seperti *term frequency-inverse document frequency* (TF-IDF).Data ini akan menjadi masukan bagi algoritma Support Vector Machine (SVM) untuk melakukan klasifikasi. TF-IDF (Term Frequency-Inverse Document Frequency) adalah teknik pembobotan dalam pengolahan teks yang digunakan untuk menilai pentingnya sebuah kata dalam sebuah dokumen relatif terhadap seluruh kumpulan dokumen (corpus). Metode ini sering digunakan dalam tugas *text mining* dan analisis data teks, termasuk analisis sentimen dan klasifikasi.

3.4. Pembagian Data

Di tahap ini, akan dilakukan pembagian data yaitu Data latih dan Data Uji.

a. Data latih adalah bagian dari dataset yang digunakan untuk melatih model SVM agar dapat mempelajari pola atau hubungan antara input (fitur) dan output (label).

b. Sedangkan, data uji adalah bagian dari dataset yang digunakan untuk mengukur kinerja model SVM setelah proses pelatihan selesai.

3.5. Support Vektor Machine

Support Vector Machine (SVM) adalah algoritma machine learning untuk klasifikasi dan regresi yang bekerja dengan mencari hyperplane terbaik untuk memisahkan data ke dalam kelas. SVM memaksimalkan margin antara hyperplane dan titik data terdekat, yang disebut support vectors. Untuk data non linear, SVM menggunakan kernel trick seperti linear, RBF, polynomial, dan sigmoid.

3.6. Naïve Bayes

Naive Bayes adalah algoritma berbasis probabilitas yang menggunakan Teorema Bayes untuk mengklasifikasikan data dengan asumsi bahwa fitur bersifat independen. Algoritma ini menghitung probabilitas posterior suatu kelas berdasarkan prior kelas dan likelihood fitur, menjadikannya cepat dan efisien untuk dataset besar. Naive Bayes sering digunakan dalam klasifikasi teks, analisis sentimen, dan deteksi spam. Meskipun efektif, kinerjanya menurun jika fitur saling bergantung atau ada data baru yang tidak dilatih.

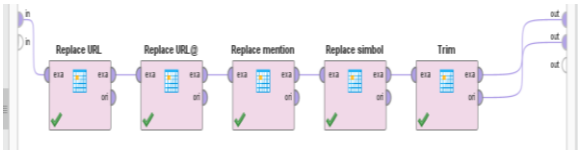
3.7. Evaluasi Model

Tahap ini menjelaskan untuk mengevaluasi kinerja dari data latih dan data uji yang terdiri dari 90:10%, 80:20%, 70:30%, 60:40%, 50:50% serta juga mengevaluasi hasil kinerja dari perbandingan klasifikasi model Support Vektor Machine dan klasifikasi model Naïve Bayes.

4. HASIL DAN PEMBAHASAN

4.1. Pengambilan Data

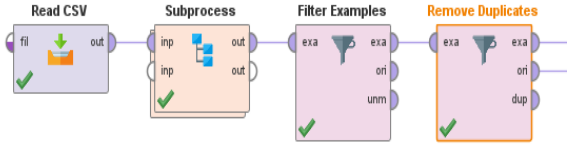
Proses crawling data menggunakan YouTube Data API v3 dilakukan untuk mengumpulkan komentar dari video di channel YouTube CNBC Indonesia terkait pembahasan mobil listrik. Data diambil dengan memasukkan API key yang telah dibuat sebelumnya dan ID video yang diambil dari URL, seperti B-WSQzz5WCg. Komentar dan balasannyadiambil menggunakan fungsi commentThreads().list dengan parameter part='snippet,replies'. Informasi yang diekstrak meliputi waktu publikasi, nama pengguna, teks komentar, dan jumlah "like". Komentar yang terkumpul diolah menggunakan library pandas untuk diorganisasi ke dalam format tabel(DataFrame) dengan kolom seperti 'publishedAt', 'authorDisplayName', 'textDisplay', dan 'likeCount'. Data tersebut kemudian disimpan dalam file CSV bernama *Data_Komen_Youtube3.csv* untuk analisis lebih lanjut. Hasil crawling berhasil mengumpulkan total 1548 komentar dari beberapa video, dengan rincian 67 komentar dari video pertama, 877 dari video kedua, dan 404 dari video ketiga. Data ini sangat terstruktur dan siap digunakan untuk analisis klasifikasi sentimen.



Gambar 2. Hasil Crawling data

4.2. Preprocessing

Tahap preprocessing bertujuan untuk mempersiapkan data teks agar sesuai untuk diproses oleh algoritma Support Vector Machine (SVM). Proses preprocessing yang dilakukan mencakup cleaningdata,casefolding,tokenization,stepword dan stemming. Alur dari proses tersebut terdapat pada gambar di bawah ini:



Gambar 3. Alur Preprocessing Data

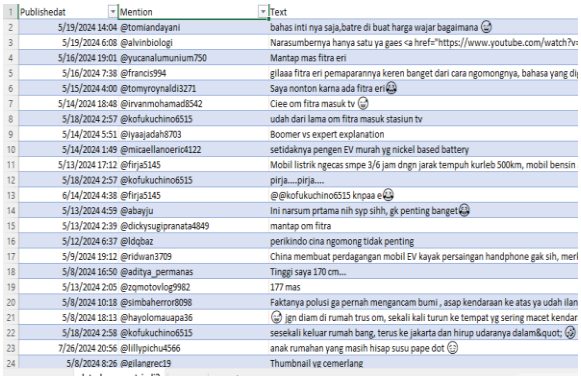
Gambar 3 tersebut menunjukkan alur proses pembersihan data yang dilakukan dengan menggunakan perangkat lunak Altair AI Studio. Proses ini dirancang untuk meningkatkan kualitas data sebelum dianalisis lebih lanjut. Berikut adalah deskripsi tiap tahapannya:

a. Read CSV: Tahap ini bertujuan untuk membaca data dalam format CSV (Comma-Separated Values) yang menjadi input utama dalam proses pembersihan. Operator ini memuat data mentah ke dalam sistem untuk diolah lebih lanjut.
b. Subprocess: Operator ini digunakan untuk menjalankan serangkaian langkah pembersihan yang telah dirancang sebelumnya dalam proses terpisah. Hal ini memungkinkan pengelompokan tugas-tugas tertentu untuk memudahkan pengelolaan alur kerja.
c. Filter Examples: Operator ini digunakan untuk menyaring data berdasarkan kriteria tertentu. Tahap ini membantu menghilangkan data yang tidak relevan atau tidak memenuhi syarat tertentu, sehingga hanya data yang sesuai yang akan diteruskan ke langkah berikutnya.
d. Remove Duplicates: Operator ini bertugas untuk mengidentifikasi dan menghapus duplikasi dalam data. Hal ini penting untuk memastikan bahwa setiap baris data yang digunakan dalam analisis adalah unik.

4.3. Cleaning Data

Pada tahap cleaning data dilakukan di dalam operator Subprocess lalu di tambah operator lain seperti Replace.Replace merupakan operasi yang digunakan untuk mengganti nilai tertentu dalam data

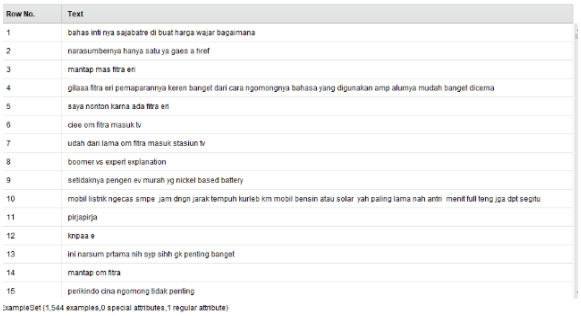
dengan nilai lain, biasanya digunakan konsisten dan sesuai dengan kebutuhan analisis.



	Publishedat	Mention	Text
2	5/19/2024 14:04	@tomiandayani	bahas inti nya saja,batre di buat harga wajar bagaimana
3	5/19/2024 6:08	@alvinbiologi	Narasumbernya hanya satu ya gaes <a href="https://www.youtube.com/watch?v=
4	5/16/2024 19:01	@yucanalumunium750	Mantap mas fitra eri
5	5/16/2024 7:38	@francis994	gilaaa fitra eri pemaparannya keren banget dari cara ngomongnya, bahasa yang di
6	5/15/2024 4:00	@tomyroynaldi3271	Saya nonton karna ada fitra eri
7	5/14/2024 18:48	@irvanmohamad8542	Ciee om fitra masuk tv
8	5/18/2024 2:57	@kofukuchino6515	udah dari lama om fitra masuk stasiun tv
9	5/14/2024 5:51	@iyaajadah8703	Boomer vs expert explanation
10	5/14/2024 1:49	@micaellanoeric4122	setidaknya pengen EV murah yg nickel based battery
11	5/13/2024 17:12	@firja5145	Mobil listrik ngecas smpe 3/6 jam dngn jarak tempuh kurleb 500km, mobil bensin
12	5/18/2024 2:57	@kofukuchino6515	pirja....pirja....
13	6/14/2024 4:38	@firja5145	@@kofukuchino6515 knpaa e
14	5/13/2024 4:59	@abayju	Ini narsum prtama nih syp sihh, gk penting banget
15	5/13/2024 2:39	@dicksugipranata4849	mantap om fitra
16	5/12/2024 6:37	@ldqbaz	perikindo cina ngomong tidak penting
17	5/9/2024 19:12	@ridwan3709	China membuat perdagangan mobil EV kayak persaingan handphone gak sih, merl
18	5/8/2024 16:50	@aditya_permanas	Tinggi saya 170 cm...
19	5/13/2024 2:05	@zqmotovlog9982	177 mas
20	5/8/2024 10:18	@simbaherror8098	Faktanya polusi ga pernah mengancam bumi , asap kendaraan ke atas ya udah ilan
21	5/8/2024 18:13	@hayolomauapa36	jgn diam di rumah trus om, sekali kali turun ke tempat yg sering macet kendar
22	5/18/2024 2:58	@kofukuchino6515	sesekali keluar rumah bang, terus ke jakarta dan hirup udaranya dalam"
23	7/26/2024 20:56	@lillypichu4566	anak rumahan yang masih hisap susu pape dot
24	5/8/2024 8:26	@gilangrec19	Thumbnail yg cemerlang

Gambar 4. Cleaning Data

Seluruh langkah dalam subprocess ini dijalankan secara berurutan untuk membersihkan data input dari elemen-elemen yang tidak relevan, seperti URL, mention, simbol, dan spasi yang tidak diperlukan. Proses ini menghasilkan output berupa teks yang telah melalui tahap pembersihan sehingga lebih terstruktur dan siap untuk digunakan pada tahap preprocessing atau analisis lebih lanjut, seperti tokenisasi dan stemming. Langkah pembersihan ini sangat penting dalam analisis data teks karena bertujuan untuk menghilangkan elemen-elemen yang dapat mengganggu kualitas hasil analisis dan memastikan bahwa data yang digunakan memiliki tingkat kebersihan dan relevansi yang optimal. Adapun hasil dari cleaning data dari gambar :



Row No.	Text
1	bahas inti nya sajabatre di buat harga wajar bagaimana
2	narasumbernya hanya satu ya gaes a href
3	mantap mas fitra eri
4	gilaaa fitra eri pemaparannya keren banget dari cara ngomongnya bahasa yang digunakan amp alurnya mudah banget dicerna
5	saya nonton karna ada fitra eri
6	ciee om fitra masuk tv
7	udah dari lama om fitra masuk stasiun tv
8	boomer vs expert explanation
9	setidaknya pengen ev murah yg nickel based battery
10	mobil listrik ngecas smpe jam dngn jarak tempuh kurleb km mobil bensin atau solar yah paling lama nah antri menit full teng jga dpt segitu
11	pirjapirja
12	knpaa e
13	ini narsum prtama nih syp sihh gk penting banget
14	mantap om fitra
15	perikindo cina ngomong tidak penting

ExampleSet (1,544 examples,0 special attributes,1 regular attribute)

Gambar 5. Hasil Cleaning Data

4.4. Labeling Data

Pada proses labeling data digunakan dengan cara labeling manual di microsof excel. Ini melibatkan memberikan label pada komentar berdasarkan ekspresi sentimen yang ada di dalamnya. Dalam prosedur ini, ulasan diberi label "positif" dan "negatif". Komentar yang menunjukkan kepuasan atau keuntungan terhadap mobil listrik tersebut diberi label "positif", sedangkan komentar yang menunjukkan ketidakpuasan atau kritik terhadap mobil listrik tersebut diberi label "negatif".Algoritma Support Vektor Machine digunakan untuk mengklasifikasikankomentar berdasarkan kata-kata atau fitur tertentu yang termasuk di dalamnya. Berdasarkan karakteristik kata-kata dalam ulasan,algoritma ini memanfaatkan probabilitas untuk memprediksi label sentimen.

Tabel 1. Data yang sudah di label

No	Text	Sentimen
1	bahas inti nya sajabatre di buat harga wajar bagaimana	negatif
2	narasumbernya hanya satu ya gaes a href	negatif
3	mantap mas fitra eri	positif
4	gilaaa fitra eri pemaparannya keren banget dari cara ngomongnya bahasa yang digunakan amp alurnya mudah banget dicerna	positif
6	saya nonton karna ada fitra eri	positif
5	ciee om fitra masuk tv	positif
6	udah dari lama om fitra masuk stasiun tv	positif
7	boomer vs expert explanation	positif
8	setidaknya pengen ev murah yg nickel based battery	negatif
….	……..	……
996	pirjapirja	negatif
997	knpaa e	negatif
998	ini narsum prtama nih syp sihh gk penting banget	negatif
999	mantap om fitra	positif
1000	perikindo cina ngomong tidak penting	negatif

4.5. Case Folding

Gambar 6. Transform Cases

Pada tahap ini case folding merupakan proses dalam pengolahan data teks yang bertujuan untuk mengubah seluruh huruf dalam teks menjadi huruf kecil (lowercase). Langkah ini dilakukan untuk menghilangkan perbedaan kapitalisasi huruf yang dapat menyebabkan kesalahan dalam analisis teks. Sebagai contoh, kata “Mobil”, “MOBIL”, dan “mobil” akan dianggap sebagai tiga entitas yang berbeda jika tidak dilakukan case folding, padahal secara semantik ketiganya merujuk pada hal yang sama.Dalam aplikasi Altair Studio atau bisa juga disebut rapid miner dengan menggunakan transform cases. Hasil dari tranform cases yaitu:

Data sebelum di Proses menggunakan tranfom cases:

Word	Attribute Name	Tota...	Docum...	negatif	positif
Mobil	Mobil	453	294	366	87
Dan	Dan	379	222	243	136
Listrik	Listrik	361	249	290	71
Di	Di	360	225	291	69
Yg	Yg	259	171	225	34
Yang	Yang	167	95	120	47
Indonesia	Indonesia	164	121	134	30

Gambar 7. data sebelum di proses

Data sesudah di Proses menggunakan tranfom cases:

Word	Attribute Name	Tota...	Docum...	negatif	positif
mobil	mobil	453	294	366	87
dan	dan	379	222	243	136
listrik	listrik	361	249	290	71
di	di	360	225	291	69
yg	yg	259	171	225	34
yang	yang	167	95	120	47
indonesia	indonesia	164	121	134	30

Gambar 8. hasil data sesudah di proses tranform cases

4.6. Tokenize

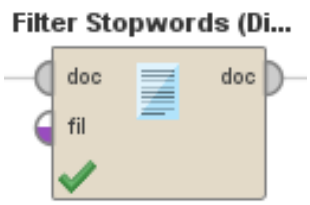
Gambar 9. Tokenize

Tokenisasi proses dasar dalam pengolahan data teks yang bertujuan untuk memecah teks menjadi unit-unit kecil yang disebut token. Token dapat berupa kata, frasa, karakter, atau elemen lain yang dianggap relevan untuk analisis. Proses ini merupakan salah satu langkah awal dalam preprocessing data teks, yang bertujuan untuk mengubah teks mentah menjadi format yang lebih terstruktur dan mudah diproses oleh algoritma atau model komputasi. Tokenisasi memecah teks menjadi bagian-bagian kecil yang disebut token, yang biasanya berupa kata atau frasa. Proses ini membantu analisis lebih lanjut karena memecah kata-kata dari teks yang panjang menjadi bagian-bagian yang lebih terstruktur dan mudah dipahami. Maka hasilnya:

Word	Attribute Name	Tota...	Docum...	negatif	positif
Mobil	Mobil	453	294	366	87
Dan	Dan	379	222	243	136
Listrik	Listrik	361	249	290	71
Di	Di	360	225	291	69
Yg	Yg	259	171	225	34
Yang	Yang	167	95	120	47
Indonesia	Indonesia	164	121	134	30
Bisa	Bisa	158	133	130	28
China	China	147	106	105	42
Itu	Itu	135	104	113	22
Negara	Negara	133	89	104	29

Gamba 10. Data hasil tokenize

4.7. Filter Stopword



Gambar 11. Filter Stopwords

Filter stopword adalah salah satu fitur dalam RapidMiner yang digunakan untuk membersihkan teks dengan menghapus kata-kata umum (stopwords) yang biasanya tidak memiliki nilai signifikan dalam analisis teks, seperti “dan,” “atau,” “dengan,” “the,” “is,” “are,” dan sebagainya. Proses ini bertujuan untuk meningkatkan efisiensi analisis dengan fokus pada kata-kata yang lebih relevan. Proses ini dimulai dengan memasukkan data teks berupa ulasan, komentar, atau dokumen ke dalam RapidMiner. Selanjutnya, daftar stopword bawaan yang tersedia untuk berbagai bahasa, termasuk bahasa Indonesia dan Inggris, dapat diterapkan untuk menyaring kata-kata tidak relevan. Proses penyaringan dilakukan dengan membandingkan setiap kata dalam teks dengan daftar stopword, lalu menghapus kata-kata yang termasuk di dalamnya. Hasil dari proses ini adalah teks yang telah difilter, yang hanya berisi kata-kata bermakna untuk langkah analisis lebih lanjut seperti stemming atau klasifikasi. Penerapan filter stopword dapat dilakukan dengan menggunakan modul Filter Stopwords yang terdapat dalam operator Text Processing pada RapidMiner, dengan langkah-langkah seperti menambahkan operator Process Documents from Data, melakukan tokenisasi, memilih daftar stopword yang sesuai, dan menjalankan proses. Hasil dari filter stopword yaitu:

Data sebelum di Proses menggunakan Filter Stopword:

Word	Attribute Name	Tota...	Docum...	negatif	positif
mobil	mobil	453	294	366	87
dan	dan	379	222	243	136
listrik	listrik	361	249	290	71
di	di	360	225	291	69
yg	yg	259	171	225	34
yang	yang	167	95	120	47
indonesia	indonesia	164	121	134	30

Gambar 12. Data sebelum di proses

Data sesudah di Proses menggunakan filter stopword:

Word	Attribut...	Tota...	Docum...	negatif	positif
mobil	mobil	453	294	366	87
listrik	listrik	361	249	290	71
yg	yg	259	171	225	34
indonesia	indonesia	164	121	134	30
china	china	147	106	105	42
negara	negara	133	89	104	29

Gambar 13. Hasil Data sesudah menggunakan filter stopword

4.8. Filter Token

Gambar 14. Filter Token

Operator “Filter Tokens (by Length)” berfungsi untuk menyaring atau membuang token (kata) yang memiliki panjang di luar batas yang ditentukan, baik terlalu pendek maupun terlalu panjang. Operator ini digunakan setelah teks melalui proses tokenisasi, di mana kalimat atau paragraf dipecah menjadi unit kata individual. Tujuan dari penyaringan ini adalah untuk menghilangkan kata-kata yang tidak bermakna secara analitis, seperti kata sangat pendek (misalnya “a”, “i”, “an”,”yg”) atau sangat panjang yang mungkin merupakan hasil kesalahan input, simbol, atau kata tidak umum. Dengan demikian, data yang dianalisis menjadi lebih bersih dan relevan, sehingga meningkatkan kualitas hasil klasifikasi atau analisis teks. Operator ini memiliki dua parameter utama, yaitu Minimum number of characters dan Maximum number of characters, yang masing-masing menentukan batas minimum dan maksimum panjang token yang diperbolehkan. Sebagai contoh, jika ditentukan minimum 3 dan maksimum 15, maka hanya kata-kata yang memiliki panjang antara 3 sampai 15 karakter saja yang akan dipertahankan, sedangkan token dengan panjang di bawah atau di atas nilai tersebut akan dihapus dari dokumen. Hasil dari filter token tersebut yaitu:

Data sebelum di Proses menggunakan filter token:

Word	Attribut...	Tota...	Docum...	negatif	positif
mobil	mobil	453	294	366	87
listrik	listrik	361	249	290	71
yg	yg	259	171	225	34
indonesia	indonesia	164	121	134	30
china	china	147	106	105	42
negara	negara	133	89	104	29
nya	nya	114	96	102	12
aja	aja	104	95	97	7
harga	harga	104	72	91	13
eropa	eropa	90	68	60	30
ya	ya	84	73	78	6
ga	ga	81	56	74	7

Gambar 15. Data sebelum di proses

Data sebelum di Proses menggunakan filter token:

Word	Attribut...	Tota...	Docum...	negatif	positif
mobil	mobil	453	294	366	87
listrik	listrik	361	249	290	71
indonesia	indonesia	164	121	134	30
china	china	147	106	105	42
negara	negara	133	89	104	29
harga	harga	104	72	91	13
eropa	eropa	90	68	60	30
murah	murah	77	70	61	16
baterai	baterai	76	42	68	8
cina	cina	74	58	61	13
produk	produk	74	49	60	14
beli	beli	70	60	60	10

Gambar 16. Hasil Data sesudah menggunakan filter token

Telihat pada hasil proses tersebut kata-kata yang tidak bermakna seperti “ya”,”ga”,”aja” dan lain nya hilang pada sesudah di proses menggunakan filter token.

4.9. Pembobotan Data

Proses pembobotan data menggunakan TF-IDF (Term Frequency-Inverse Document Frequency) bertujuan untuk mentransformasi data teks hasil praproses menjadi representasi numerik yang siap digunakan oleh algoritma pembelajaran mesin seperti SVM. Data awal diimpor dari file CSV menggunakan operator ReadCSV, kemudian atribut nominal diubah menjadi teks melalui Nominal_toText. Selanjutnya, data teks diolah dengan operator Process_DocumentsFromData, yang meliputi langkah-langkah seperti normalisasi teks, tokenisasi, penghapusan *stopwords*, dan *stemming*. Hasil pengolahan ini diubah menjadi data terstruktur menggunakan WordList_toData, diurutkan dengan operator Sort, dan disaring menggunakan Filter_ExampleRange untuk memilih subset data yang relevan. Pada tahap pengolahan teks di RapidMiner, operator seperti Tokenize, Transform_Cases, Filter_Stopwords, dan Filter_Tokens digunakan untuk membersihkan serta menyaring data teks. Proses ini diakhiri dengan pembuatan vektor menggunakan TF-IDF, yang memberikan bobot lebih tinggi pada kata-kata signifikan yang sering muncul di dokumen tertentu tetapi jarang ditemukan di korpus keseluruhan, sehingga meningkatkan relevansi fitur untuk analisis sentimen.

Row No.	word	in documents	total	in class (NEGATIVE)	in class (POSITIVE)
1	mobil	294	453	451	2
2	listrik	249	361	360	1
3	indonesia	121	164	160	4
4	china	106	147	144	3
5	negara	89	133	132	1
6	harga	72	104	104	0
7	eropa	68	90	89	1
8	murah	70	77	77	0
9	baterai	42	76	76	0
10	cina	58	74	73	1
11	produk	49	74	74	0
12	beli	60	70	69	1
13	amerika	41	66	65	1
14	barat	42	61	61	0
15	kendaraan	45	60	59	1

ExampleSet (1,000 examples, 0 special attributes, 5 regular attributes)

Gambar 17. Hasil TF-IDF

4.10. Pembagian Data

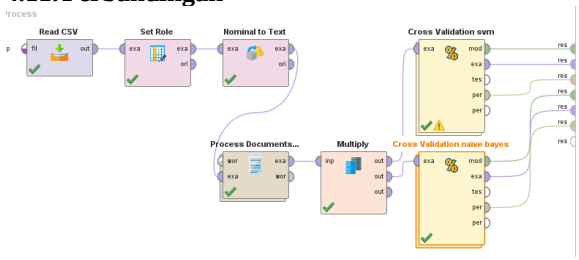
Tabel 2. Pembagian Dataset

No.	Model	Data Latih	Data Uji
1.	Model Pertama	90%	10%
2.	Model Kedua	80%	20%
3.	Model Ketiga	70%	30%
4.	Model Keempat	60%	40%
5.	Model Kelima	50%	50%

Tabel tersebut menampilkan pemisahan data pengujian dan pelatihan dalam pengembangan m odel klasifikasi sentimen. Kedua set data ini sangat

penting dalam proses pembelajaran mesin karena set pelatihan digunakan untuk melatih model, dan set pengujian digunakan untuk membandingkan kinerja model dengan data yang sebelumnya tidak terlihat. Lima skenario berbeda membentuk pembagian data dalam tabel, dan masing-masing skenario menggambarkan berbagai rasio pelatihan terhadap data uji.

4.11. Perbandingan



Gambar 18. Perbandingan SVM dan Naive Bayes

Gambar 18 menjelaskan perbandingan algoritma Support Vector Machine (SVM) dan Naïve Bayes dalam klasifikasi teks menggunaka Cross Validation. Tahapan dimulai dari pembacaan dataset (CSV), dilanjutkan dengan penentuan peran atribut sebagai fitur atau label. Data nominal dikonversi menjadi teks untuk mempermudah pemrosesan seperti tokenisasi, penghapusan stopwords, dan stemming. Selanjutnya, data diproses menggunakan kedua algoritma untuk perbandingan performa. Evaluasi dilakukan dengan Cross Validation, misalnya 10-fold, menggunakan metrik seperti akurasi, presisi, recall, dan F1-score. Hasil evaluasi memberikan gambaran efektivitas masing-masing algoritma dalam menangani klasifikasi teks.

4.12. Hasil akurasi,presisi ,recall dan F1-score

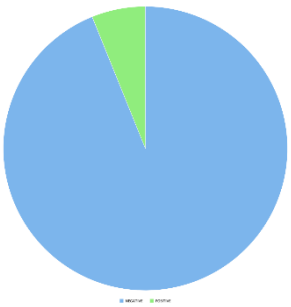
Tahapan evaluasi bertujuan menilai kinerja model klasifikasi sentimen dengan metrik seperti akurasi, presisi, recall, dan F1-score, yang dievaluasi pada dataset uji. Metrik ini memberikan gambaran seberapa baik model mengklasifikasikan sentimen komentar, baik positif maupun negatif, sehingga dapat memahami persepsi publik terhadap mobil listrik. Selain itu, matriks konfusi digunakan untuk merepresentasikan prediksi model yang benar dan salah. Sebelum evaluasi, model dilatih pada data latih yang dilabeli sentimen melalui proses ekstraksi fitur dan pembelajaran pola menggunakan algoritma seperti SVM. Evaluasi ini memastikan model mampu mengenali pola sentimen secara akurat sebelum diterapkan pada data baru.

Tabel 3. Hasil hasil akurasi, presisi, recall, dan F1-score dari berbagai pembagian data

Data Split	Accuracy (%)	Error (%)	Prec. Neg (%)	Prec. Pos (%)	Rec. Neg (%)	Rec. Pos (%)	F1 Neg (%)	F1 Pos (%)
90%-10%	93.88	0.58	91.88	0	100	0	91.88	0
80%-20%	93.87	0.18	91.87	0	100	0	91.87	0
70%-30%	93.86	0.09	93.86	0	100	0	93.86	0
60%-40%	93.80	0.63	93.80	0	100	0	93.80	0

4.13. Hasil WourCloud

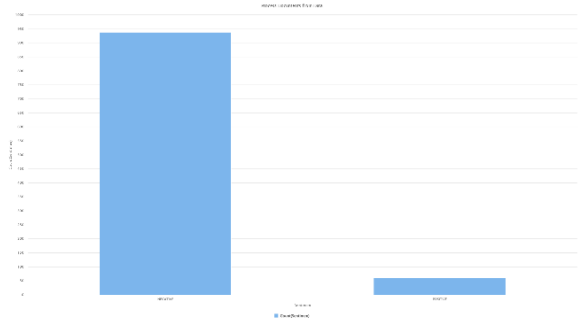
a. Piechart



Gambar 19. Piechart Negatif dan Poitif

Diagram pie di atas menggambarkan distribusi data berdasarkan dua kelas, yaitu negative dan positive. Sebagian besar data termasuk dalam kelas negative, yang ditandai dengan warna biru, sementara kelas positive, yang ditandai dengan warna hijau, hanya mewakili bagian kecil dari keseluruhan data. Ketidakseimbangan yang signifikan ini mencerminkan adanya class imbalance dalam dataset, di mana jumlah data pada kelas negative jauh lebih besar dibandingkan kelas positive.

b. BarColoumn



Gambar 20. Bar Coloumn Negatif dan Posistif

Diagram batang di atas menggambarkan distribusi jumlah data berdasarkan dua kategori sentimen, yaitu negative dan positive. Kategori negative menunjukkan batang yang sangat tinggi, yang menandakan bahwa mayoritas data dalam dataset termasuk dalam sentimen negatif. Sebaliknya, kategori positive memiliki batang yang jauh lebih pendek, menunjukkan bahwa jumlah data dengan sentimen positif sangat sedikit.

4.14. Hasil Perbandingan SVM dan Naïve Bayes

Hasil dari perbanding algoritma Support Vektor Machine dan Naïve Bayes cukup ada perbedaan antara True Positif (TP), False Positif (FP), True Negatifr(TN), dan False Negatif (FN).Berikut adalah gambar hasil dari perbandingan Support Vektor Machine dan Naïve Bayes:

Algoritma	Kelas	True Negative (TN)	False Positive (FP)	True Positive (TP)	False Negative (FN)	Presisi (%)	Recall (%)
Naïve Bayes	Negative	899	30	-	-	96.77	95.84
	Positive	-	-	31	39	44.29	50.82
SVM	Negative	938	0	-	-	93.89	100
	Positive	-	-	0	70	0	0

Gambar 21. Hasil Perbandingan SVM dan Naïve Bayes

4.15. Naive Bayes

a. Kelas Negative
   - True negative:dari 929 data sebenarnya negative, sebanyak 899 berhasil diprediksi dengan benar sebagai negative.
   - False positive:sebanyak 30 data yang sebenarnya negative salah diklasifikasikan sebagai positive.
   - Evaluasi:naive bayes memiliki presisi untuk kelas negative sebesar 96.77%, artinya sebagian besar prediksi negative benar. Recall untuk kelas negative sebesar 95.84%, menunjukkan hampir semua data negative dikenali dengan benar.

b. Kelas Positive
   - True positive:dari 70 data sebenarnya positiv e, hanya 31 yang diprediksi dengan benar sebagai positive.
   - False negative:sebanyak 39 data positive sal ah diprediksi sebagai negative.
   - Evaluasi:presisi untuk kelas positive hanya 44.29%, menunjukkan banyak kesalahan prediksi positive. Recall untuk kelas positive adalah 50.82%, artinya hanya sekitar separuh data positive dikenali dengan benar.

4.16. Support Vector Machine (SVM)

a. Kelas Negative
   - True negative:dari 938 data sebenarnya nega tive, semuanya diprediksi dengan benar seba gai negative (true negative: 938).
   - False positive:tidak ada kesalahan klasifikasi negative sebagai positive (false positive: 0).
   - Evaluasi:presisi kelas negative adalah 93.89%, sedikit lebih rendah dari naive bayes, tetapi sangat andal. Recall untuk kelas negative adalah 100%, menunjukkan svm mengenali semua data negative dengan sempurna.

b. Kelas Positive
   - True positive:tidak ada data sebenarnya positive yang berhasil diprediksi dengan benar (true positive: 0).
   - False negative:semua data sebenarnya positi ve salah diklasifikasikan sebagai negative (false negative: 70).
   - Evaluasi:presisi kelas positive adalah 0%, karena tidak ada data positive yang diprediksi dengan benar. Recall untuk kelas positive juga 0%, artinya model sepenuhnya gagal mengenali kelas ini.

5. KESIMPULAN DAN SARAN

Hasil penelitian menunjukkan bahwa algoritma Support Vector Machine (SVM) memiliki kinerja yang baik dalam mengklasifikasikan komentar negatif, dengan 938 komentar negatif berhasil diklasifikasikan dengan benar, meskipun masih terdapat 61 komentar positif yang salah diklasifikasikan sebagai negatif (false positive), mengindikasikan adanya bias dalam mendeteksi komentar positif. Akurasi model SVM berkisar antara 93,6% hingga 93,88%, dengan akurasi tertinggi pada skenario pembagian data 90:10, yang menunjukkan bahwa proporsi data pelatihan yang lebih besar memberikan hasil yang lebih optimal, sedangkan penurunan akurasi terjadi pada skenario dengan data latih lebih sedikit. SVM juga menunjukkan akurasi dan kestabilan yang lebih baik dibandingkan Naïve Bayes (93,89% ± 0,56% vs. 93,09% ± 2,13%), namun SVM memiliki kelemahan signifikan dalam mengenali sentimen positif dengan presisi dan recall sebesar 0,00%. Oleh karena itu, disarankan untuk mengatasi bias terhadap kelas negatif dengan teknik penyeimbangan kelas seperti SMOTE atau undersampling, melakukan eksperimen dengan pembagian data latih dan uji yang beragam menggunakan cross-validation, serta menerapkan teknik ensemble seperti bagging atau boosting yang menggabungkan keunggulan SVM dan Naïve Bayes guna meningkatkan presisi dan recall pada kedua kelas sentimen dan menghasilkan model klasifikasi yang lebih seimbang dan akurat.

DAFTAR PUSTAKA

[1] P. Arsi and R. Waluyo, "Analisis Sentimen Wacana Pemindahan Ibu Kota Indonesia Menggunakan Algoritma Support Vector Machine (SVM)," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 8, no. 1, p. 147, 2021, doi: 10.25126/jtiik.0813944.

[2] P. B. Bahtera and D. S. Kartawijaya, "Content Classification of the Official Website of the Ministry of Foreign Affairs of the Republic of Indonesia (MoFA RI) using Vector Space Model (VSM)," *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 4, no. 4, pp. 1309–1319, 2024, doi: 10.57152/malcom.v4i4.1368.

[3] M. Deori, V. Kumar, and M. K. Verma, "Analysis of YouTube video contents on Koha

and DSpace, and sentiment analysis of viewers' comments," *Libr. Hi Tech*, vol. 41, no. 3, pp. 711–728, 2023, doi: 10.1108/LHT-12-2020-0323.

[4] E. Ditendra, S. Suryani, S. Romelah, M. H. Arsyiddik Tanjung, and M. Sarah, "Perbandingan Algoritma Klasifikasi untuk Analisis Sentimen Islam Nusantara di Indonesia," *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 2, no. 1, pp. 71–77, 2022, doi: 10.57152/malcom.v2i1.199.

[5] J. Homepage, D. Pramudita, Y. Akbar, and T. Wahyudi, "MALCOM: Indonesian Journal of Machine Learning and Computer Science Sentiment Analysis of the Indonesian Smart College Card Program on Social Media X Using the Naive Bayes Algorithm Analisis Sentimen Terhadap Program Kartu Indonesia Pintar Kuliah Pada Med," *Malcom* , vol. 4, no. October, pp. 1420–1430, 2024.

[6] Merinda Lestandy, Abdurrahim Abdurrahim, and Lailis Syafa'ah, "Analisis Sentimen Tweet Vaksin COVID-19 Menggunakan Recurrent Neural Network dan Naïve Bayes," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 5, no. 4, pp. 802–808, 2021, doi: 10.29207/resti.v5i4.3308.

[7] D. Musfiroh, U. Khaira, P. E. P. Utomo, and T. Suratno, "Analisis Sentimen terhadap Perkuliahan Daring di Indonesia dari Twitter Dataset Menggunakan InSet Lexicon," *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 1, no. 1, pp. 24–33, 2021, doi: 10.57152/malcom.v1i1.20.

[8] K. S. Parikh and T. P. Shah, "Support Vector Machine – A Large Margin Classifier to Diagnose Skin Illnesses," *Procedia Technol.*, vol. 23, pp. 369–375, 2016, doi: 10.1016/j.protcy.2016.03.039.

[9] T. M. Permata Aulia, N. Arifin, and R. Mayasari, "Perbandingan Kernel Support Vector Machine (Svm) Dalam Penerapan Analisis Sentimen Vaksinisasi Covid-19," *SINTECH (Science Inf. Technol. J.*, vol. 4, no. 2, pp. 139–145, 2021, doi: 10.31598/sintechjournal.v4i2.762.

[10] R. S. Putra and I. D. Ratih, "Klasifikasi Tanggapan Pelaksanaan Program Magang dengan Menggunakan Metode Naive Bayes Classifier," *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 1, no. 2, pp. 129–137, 2021, doi: 10.57152/malcom.v1i2.113.

[11] S. Rabbani, D. Safitri, N. Rahmadhani, A. A. F. Sani, and M. K. Anam, "Perbandingan Evaluasi Kernel SVM untuk Klasifikasi Sentimen dalam Analisis Kenaikan Harga BBM," *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 3, no. 2, pp. 153–160, 2023, doi: 10.57152/malcom.v3i2.897.

[12] A. P. Rodrigues *et al.*, "Real-Time Twitter Spam Detection and Sentiment Analysis using Machine Learning and Deep Learning Techniques," *Comput. Intell. Neurosci.*, vol. 2022, 2022, doi: 10.1155/2022/5211949.

[13] D. Sari and R. Kurniawan, "Sentiment Analysis of The Performance of Medan Mayor Program on Social Media X Using Support Vector Machine Analisis Sentimen Terhadap Kinerja Program Walikota Medan pada Media Sosial X Menggunakan Support Vector Machine," vol. 4, no. October, pp. 1539–1548, 2024.

[14] M. Wankhade, A. C. S. Rao, and C. Kulkarni, *A survey on sentiment analysis methods, applications, and challenges*, vol. 55, no. 7. Springer Netherlands, 2022. doi: 10.1007/s10462-022-10144-1.

[15] R. S. Wijaya, A. Qur'ania, and I. Anggraeni, "Klasifikasi Penyakit Cacar Monyet Menggunakan Support Vector Machine (SVM)," *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 4, no. 4, pp. 1253–1260, 2024, doi: 10.57152/malcom.v4i4.1417