INOTEK, Vol. 9
ISSN: 2580-3336 (Print) / 2549-7952 (Online)
Url: https://proceeding.unpkediri.ac.id/index.php/inotek/

Analisis Sentimen Komentar Video Youtube Dengan Leksikon Textblob Menggunakan Algoritma Svm Berdasarkan Rasio Data Uji

[1*]Fredi Wijaya, [2]Risky Aswi Ramadhani, [3]Ahmad Bagus Setiawan
[1 2 3] Teknik Informatika, Universitas Nusantara PGRI Kediri
E-mail: *[1]*frediw14@gmail.com*, [2]*riskyaswiramadhani@gmail.com*, [3]*ahmadbagus@gmail.com*
Penulis Korespondens : Risky Aswi Ramadhani

Abstrak — Analisis sentimen, sebagai salah satu cabang penting dalam *Natural Language Processing* (NLP), bertujuan untuk mengidentifikasi dan memahami opini atau emosi yang tersirat dalam data teks. Penelitian ini secara spesifik berfokus pada analisis sentimen komentar pengguna yang terdapat pada satu video YouTube yang mengulas produk laptop *Advan WorkPlus*. Tujuan utama dari penelitian ini adalah untuk mengevaluasi secara komparatif performa akurasi model SVM dengan menerapkan variasi pada rasio pembagian data latih dan data uji. Tiga rasio yang diuji meliputi data uji sebesar 10%, 20%, dan 30%. Akurasi tertinggi yang berhasil dicapai adalah 72,00% ketika menggunakan rasio data uji 10%. Terlihat pola penurunan akurasi seiring dengan peningkatan persentase data uji, di mana akurasi menjadi 66,00% untuk data uji 20% dan 64,00% untuk data uji 30%. Pengujian ini konsisten mengindikasikan bahwa ketersediaan pengujian data latih yang lebih besar memiliki korelasi positif terhadap peningkatan performa akurasi model dalam analisis sentimen yang memanfaatkan *TextBlob* dan SVM.
Kata Kunci— Akurasi, Analisis Sentimen, Rasio Data Uji, SVM, TextBlob, *YouTube*

Abstract *— Sentiment analysis, as one of the important branches in Natural Language Processing (NLP), aims to identify and understand opinions or emotions implied in text data. This study specifically focuses on the sentiment analysis of user comments contained in a YouTube video reviewing the Advan WorkPlus laptop product. The main objective of this study is to comparatively evaluate the accuracy performance of the SVM model by applying variations in the ratio of training data and test data sharing. The three ratios tested include test data of 10%, 20%, and 30%. The highest accuracy achieved was 72.00% when using a test data ratio of 10%. A pattern of decreasing accuracy can be seen along with the increase in the percentage of test data, where the accuracy becomes 66.00% for 20% test data and 64.00% for 30% test data. This test consistently indicates that the availability of larger training data testing has a positive correlation with the increase in model accuracy performance in sentiment analysis utilizing TextBlob and SVM.*
Keywords*— Accuracy, Sentiment Analysis, Test Data Ratio, SVM, TextBlob,YouTube*

I. PENDAHULUAN

Pesatnya adopsi internet dan platform media sosial telah membentuk lanskap komunikasi digital yang baru, di mana jutaan pengguna setiap hari berbagi pikiran, pengalaman, dan opini mereka. *YouTube*, sebagai platform berbagi video terbesar di dunia, bukan hanya menjadi tempat untuk mengunggah dan menonton video, tetapi juga sebagai forum diskusi aktif melalui kolom komentarnya. Komentar yang diunggah oleh pengguna di *YouTube* seringkali mengandung sentimen tersirat yang sangat berharga, yang jika dianalisis, dapat memberikan wawasan mendalam mengenai persepsi publik terhadap berbagai topik, mulai dari berita, isu sosial, hingga produk dan layanan [1].

Analisis sentimen, yang juga dikenal sebagai opinion mining, adalah cabang *Natural Language Processing* (NLP) yang memiliki tujuan fundamental untuk secara otomatis mengidentifikasi, mengekstrak, dan mengklasifikasikan polaritas sentimen (misalnya, positif, negatif, atau netral) dari suatu teks [5]. Dalam konteks komersial, kemampuan untuk memahami sentimen konsumen secara *real-time* terhadap suatu produk menjadi krusial. Misalnya, bagi produsen laptop, komentar pengguna di *YouTube* mengenai produk mereka, seperti laptop *Advan WorkPlus*, dapat menjadi umpan balik yang tidak ternilai untuk memahami kekuatan dan kelemahan produk di mata konsumen, serta untuk merumuskan strategi pemasaran dan pengembangan produk di masa mendatang [2].

Sebelum dilakukan pelabelan sentimen, komentar-komentar ini menjalani tahapan pra-pemrosesan yang penting, seperti case folding, tokenisasi, penghapusan stop word, dan stemming untuk membersihkan serta menstandarkan data. Proses ini penting untuk meminimalkan *noise* dan menyiapkan teks agar dapat dianalisis dengan efektif oleh model.. *TextBlob* adalah pustaka *Python* yang menyediakan antarmuka sederhana untuk melakukan tugas-tugas NLP umum, termasuk analisis sentimen berbasis leksikon. Leksikon *TextBlob* bekerja dengan memberikan skor polaritas (derajat positif atau negatif) dan subjektivitas pada kata-kata, yang kemudian digabungkan untuk memberikan skor sentimen pada kalimat atau dokumen secara keseluruhan [6]. Kelebihan *TextBlob* terletak pada kemudahan penggunaannya, meskipun untuk bahasa selain Inggris mungkin memerlukan adaptasi atau translasi data [7].

Algoritma Support Vector Machine (SVM) diterapkan dalam penelitian ini untuk melakukan klasifikasi sentimen pada komentar-komentar yang telah diproses dan dibobotkan. SVM digunakan untuk mengidentifikasi dan memisahkan komentar ke dalam kategori sentimen positif, negatif, atau netral dengan mencari hyperplane optimal yang memaksimalkan margin antar kelas[3]. Keunggulan SVM terletak pada kemampuannya menangani data berdimensi tinggi dan generalisasi yang baik, bahkan dengan jumlah pengujian data latih yang relatif terbatas [4].

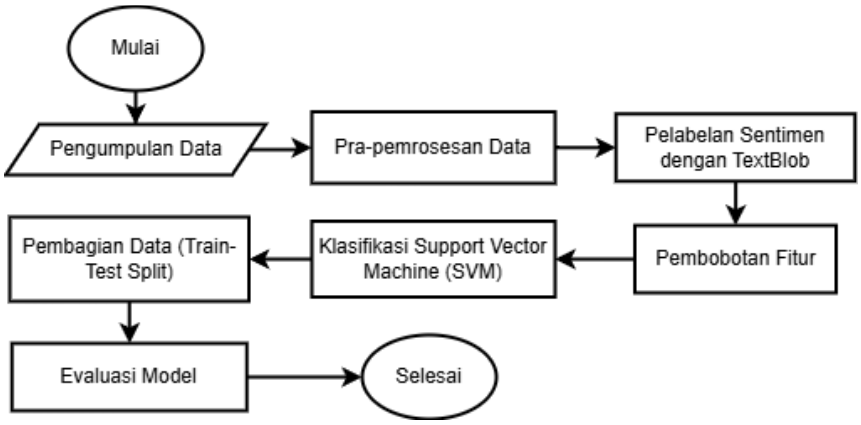
Penelitian ini memiliki fokus yang spesifik pada analisis sentimen komentar dari satu video *YouTube*, yaitu video David GadgetIn dengan Video ID ‘gp8JOmHEfos’ yang mengulas laptop *Advan WorkPlus*. Berbeda dengan penelitian sebelumnya yang mungkin melibatkan berbagai sumber data atau leksikon, penelitian ini sengaja membatasi ruang lingkup pada satu video dan penggunaan hanya leksikon *TextBlob*. Tujuan utamanya adalah untuk membandingkan dan menganalisis performa akurasi model klasifikasi sentimen berbasis SVM ketika rasio pembagian data latih dan data uji divariasikan. Tiga skenario rasio data uji yang akan dieksplorasi adalah 10%, 20%, dan 30%. Dengan demikian, penelitian ini berupaya menjawab rumusan masalah inti: "Bagaimana akurasi analisis sentimen komentar video YouTube yang spesifik ini menggunakan leksikon *TextBlob* dan algoritma SVM ketika data uji divariasikan sebesar 10%, 20%, dan 30%". Diharapkan hasil penelitian ini dapat memberikan wawasan

mengenai dampak ukuran data latih terhadap akurasi model analisis sentimen dalam konteks aplikasi yang terfokus.

Penelitian ini menganalisis sentimen komentar dari satu video *YouTube* David GadgetIn dengan Video ID 'gp8JomHEfos' yang membahas review laptop *Advan WorkPlus*. Fokusnya adalah membandingkan akurasi model SVM dengan leksikon *TextBlob* pada variasi rasio data uji: 10%, 20%, dan 30%. Rumusan masalahnya adalah bagaimana akurasi model pada setiap rasio data uji.

II. METODE

Metode penelitian ini mencakup beberapa tahapan.



Gambar 1. Alur Pemrosesan

2.1 Pengumpulan Data

Komentar dikumpulkan dari satu video YouTube David GadgetIn (Video ID: gp8JOmHEfos) yang mengulas laptop *Advan WorkPlus*, menggunakan YouTube Data API v3. Berikut adalah *WordCloud* dari 500 data yang telah diambil.

Gambar 2. WordCloud Data

2.2 Pra-pemrosesan Data

Tahapan ini meliputi: *Cleansing* (Menghapus karakter-karakter yang tidak relevan atau *noise* dari teks, seperti URL, tag HTML, atau simbol-simbol yang tidak bermakna), *Case Folding* (Semua huruf dalam teks komentar diubah menjadi huruf kecil), *Tokenisasi* (Proses memecah atau membagi aliran teks menjadi unit-unit yang lebih kecil yang disebut token), *Stop Word Removal* (Menghapus kata tidak berbobot), dan *Stemming* (pengubahan kata ke bentuk dasar) [8].

Tabel 1. Cleansing

Sebelum Cleansing	Sesudah Cleansing
Adeku saranin jangan brand lokal ataupun china :(Padahal ini keknya oke	Adeku saranin jangan brand lokal ataupun china Padahal ini keknya oke
Harga segini untuk lokal sangat OKKE 1:21	Harga segini untuk lokal sangat OKKE
Kenapa ya kok port USB dikit, dan LAN tuh dihilangkan? Laptop gen now rata2 begini	Kenapa ya kok port USB dikit dan LAN tuh dihilangkan Laptop gen now rata begini

Tabel 2. Case Folding

Cleansing	Case Folding
Adeku saranin jangan brand lokal ataupun china Padahal ini keknya oke	adeku saranin jangan brand lokal ataupun china padahal ini keknya oke
Harga segini untuk lokal sangat OKKE	harga segini untuk lokal sangat okke
Kenapa ya kok port USB dikit dan LAN tuh dihilangkan Laptop gen now rata begini	kenapa ya kok port usb dikit dan lan tuh dihilangkan laptop gen now rata begini

Tabel 3. Tokenisasi

Case Folding	Tokenisasi
adeku saranin jangan brand lokal ataupun china padahal ini keknya oke	['adeku', 'saranin', 'jangan', 'brand', 'lokal', 'ataupun', 'china', 'padahal', 'ini', 'keknya', 'oke']
harga segini untuk lokal sangat okke	['harga', 'segini', 'untuk', 'lokal', 'sangat', 'okke']
kenapa ya kok port usb dikit dan lan tuh dihilangkan laptop gen now rata begini	['kenapa', 'ya', 'kok', 'port', 'usb', 'dikit', 'dan', 'lan', 'tuh', 'dihilangkan', 'laptop', 'gen', 'now', 'rata', 'begini']

Tabel 4. Stopword

Tokenisasi	Stopword
['adeku', 'saranin', 'jangan', 'brand', 'lokal', 'ataupun', 'china', 'padahal', 'ini', 'keknya', 'oke']	['adeku', 'saranin', 'brand', 'lokal', 'china', 'keknya', 'oke']
['harga', 'segini', 'untuk', 'lokal', 'sangat', 'okke']	['harga', 'segini', 'lokal', 'okke']
['kenapa', 'ya', 'kok', 'port', 'usb', 'dikit', 'dan', 'lan', 'tuh', 'dihilangkan', 'laptop', 'gen', 'now', 'rata', 'begini']	['ya', 'port', 'usb', 'dikit', 'lan', 'tuh', 'dihilangkan', 'laptop', 'gen', 'now']

Tabel 5. Stemming

Stopword	Stemming
['adeku', 'saranin', 'brand', 'lokal', 'china', 'keknya', 'oke']	['adeku', 'saranin', 'brand', 'lokal', 'china', 'kek', 'oke']
['harga', 'segini', 'lokal', 'okke']	['harga', 'gin', 'lokal', 'okke']
['ya', 'port', 'usb', 'dikit', 'lan', 'tuh', 'dihilangkan', 'laptop', 'gen', 'now']	['ya', 'port', 'usb', 'dikit', 'lan', 'tuh', 'hilang', 'laptop', 'gen', 'now']

2.3 Pelabelan Sentimen dengan TextBlob

Setelah data melewati proses preprocessing, kemudian data akan di translate ke Bahsa Inggris dan setiap komentar diberi label sentimen menggunakan leksikon TextBlob. Polaritas TextBlob (-1.0 hingga +1.0) digunakan: >0 (Positif), <0 (Negatif), =0 (Netral) [9].

Tabel 6. score leksikon TextBlob

Data Preprocessing	Translate Data	Score Leksikon TextBlob
adeku saranin brand lokal china kek oke	advice adviceine local brand china kek okay	0.25
harga gin lokal okke	local gin price okke	0.0
a port usb dikit lan tuh hilang laptop gen now	Yes, the USB port a little lan is lost a laptop gen now	-0.1875

2.4 Pembobotan Fitur

Teks yang sudah bersih dan memiliki label sentimen harus diubah menjadi format numerik yang dapat dipahami oleh algoritma machine learning. Metode yang digunakan untuk pembobotan fitur adalah Term Frequency-Inverse Document Frequency (TF-IDF). TF-IDF adalah teknik statistik yang banyak digunakan dalam NLP untuk mengukur seberapa penting sebuah kata dalam dokumen dalam sebuah korpus. Semakin tinggi nilai TF-IDF suatu kata, semakin relevan kata tersebut bagi dokumen tertentu dalam koleksi dokumen [3].

Tabel 7. Tf-Idf

Kata	Bobot
advan	34.01961898051793
laptop	24.867536856260713
beli	20.44349170114667

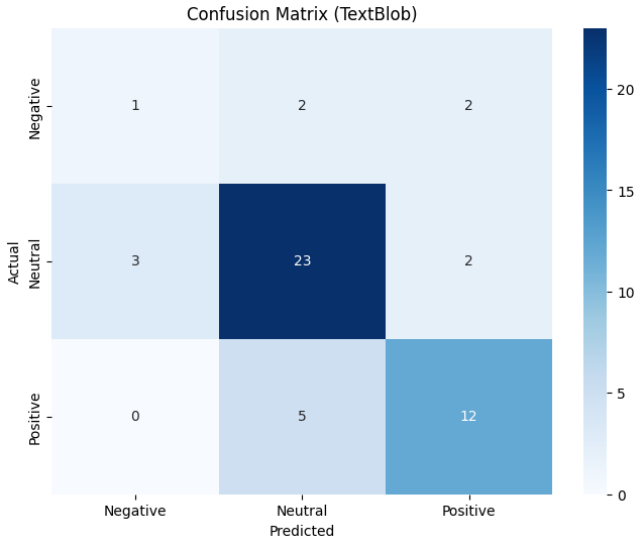
2.5 Klasifikasi dengan Support Vector Machine (SVM)

Data yang telah diubah menjadi representasi fitur numerik kemudian digunakan untuk melatih dan menguji model klasifikasi. Algoritma Support Vector Machine (SVM) dipilih karena efektivitasnya yang telah terbukti dalam berbagai tugas klasifikasi teks. SVM bekerja dengan mencari hyperplane optimal yang memaksimalkan margin pemisah antara kelas-kelas yang berbeda. Untuk data yang tidak dapat dipisahkan secara linear, SVM menggunakan fungsi kernel untuk memetakan data ke ruang dimensi yang lebih tinggi di mana pemisahan linear dimungkinkan [10].

2.6 Pembagian Data (Train-Test Split)

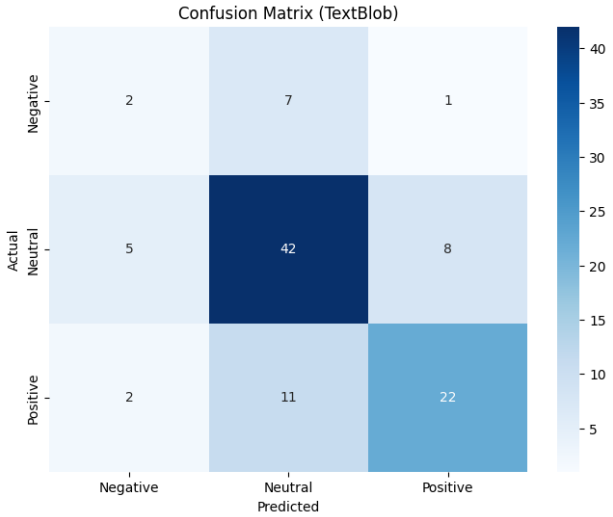
Tiga variasi rasio data diuji:

1. Rasio 1: Data Latih 90%, Data Pengujian 10%



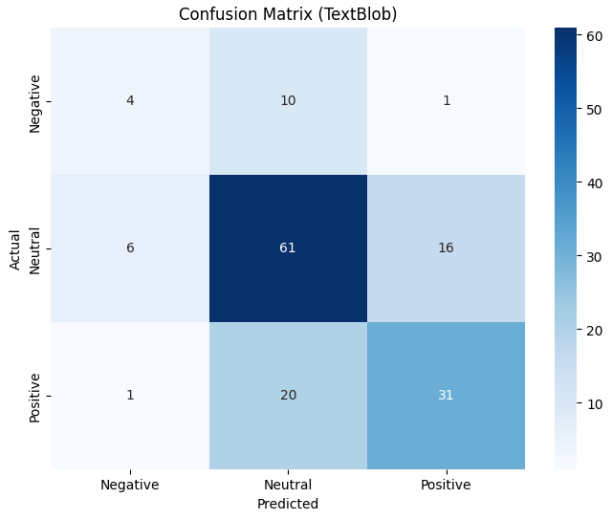
Gambar 3. Heatmap 10%

2. Rasio 2: Data Latih 80%, Data Pengujian 20%



Gambar 4. Heatmap 20%

3. Rasio 3: Data Latih 70%, Data Pengujian 30%



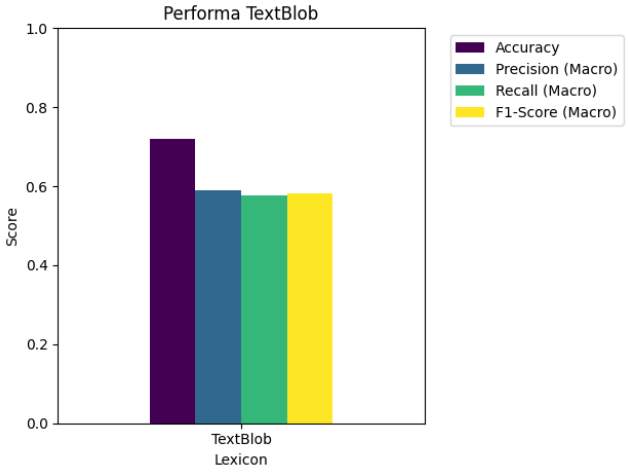
Gambar 5. Heatmap 30%

2.7 Evaluasi Model

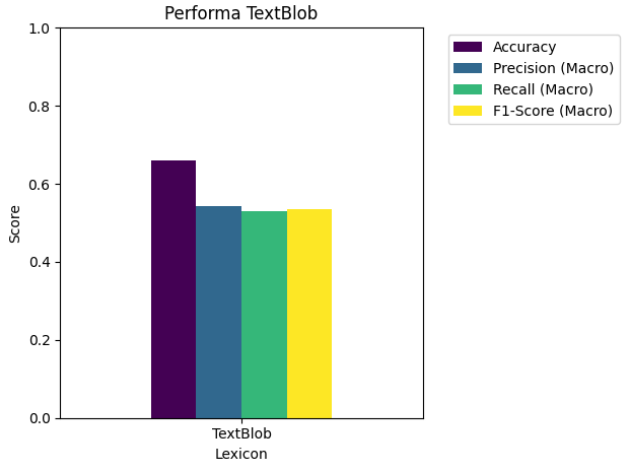
Kinerja model diukur dengan metrik akurasi:

$$Akurasi = \frac{Jumlah\ Prediksi\ Benar}{Jumlah\ Total\ Prediksi} \quad (1)$$

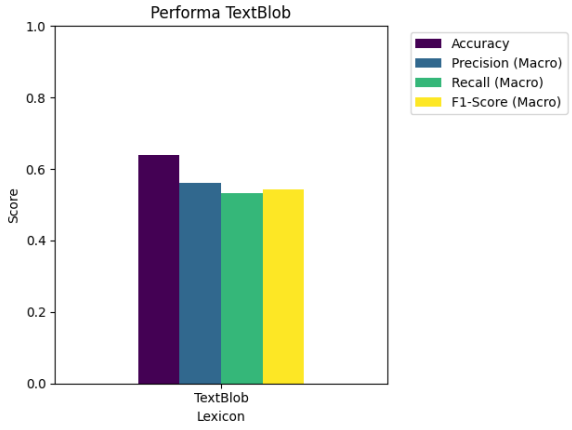
Akurasi dari data testign 10% = 0.7200 = 72%
Akurasi dari data testign 20% = 0.6600 = 66%
Akurasi dari data testign 30% = 0.6400 = 64%



Gambar 6. Performa 10% Data Testing



Gambar 7. Performa 20% Data Testing



Gambar 8. Performa 30% Data Testing

III. HASIL DAN PEMBAHASAN

Tren yang diamati dari hasil pengujian ini menunjukkan hubungan yang konsisten dan dapat diprediksi yaitu semakin kecil proporsi data yang dialokasikan untuk pengujian, semakin tinggi

pula akurasi yang mampu dicapai oleh model klasifikasi. Fenomena ini dapat dijelaskan secara teoritis dalam konteks *machine learning*. Dengan lebih banyak data latih, model SVM memiliki kesempatan yang lebih luas untuk mempelajari berbagai pola yang mendasari dan karakteristik sentimen dari data komentar. Hal ini memungkinkan model untuk membangun *hyperplane* pemisah yang lebih robust dan optimal, yang pada gilirannya meningkatkan kemampuan model untuk menggeneralisasi dengan akurat terhadap data yang belum pernah dilihat sebelumnya atau data uji**.**

Penggunaan leksikon *TextBlob*, meskipun tidak secara spesifik dioptimalkan untuk karakteristik dan nuansa bahasa Indonesia, masih mampu menghasilkan tingkat akurasi yang cukup menjanjikan, terutama pada rasio data uji 10%. Ini menunjukkan bahwa *TextBlob* memiliki kapabilitas dasar yang dapat dimanfaatkan untuk analisis sentimen pada teks berbahasa Indonesia, meskipun mungkin terdapat keterbatasan dalam menangani ekspresi idiomatik, bahasa gaul, atau konteks budaya lokal yang tidak sepenuhnya tercakup dalam kamus leksikonnya. Namun, hasil ini juga membuka peluang untuk penelitian lebih lanjut dalam pengembangan leksikon sentimen khusus bahasa Indonesia yang lebih adaptif.

Penurunan akurasi yang terjadi seiring dengan peningkatan ukuran data uji menggarisbawahi pentingnya keseimbangan antara data latih dan data uji. Ketika porsi data yang dilatih berkurang secara signifikan (misalnya dari 90% menjadi 70%), model memiliki lebih sedikit contoh untuk belajar, sehingga kemampuan generalisasinya dapat terganggu. Ini menyebabkan performa model pada data baru (data uji) menjadi kurang akurat. Dalam studi ini, perbedaan akurasi antara rasio data uji 10% dan 20% tidak terlalu mencolok, namun penurunan menjadi lebih signifikan saat data uji mencapai 30%. Hal ini mungkin mengindikasikan bahwa untuk dataset komentar *YouTube* yang digunakan, titik optimal atau cukup dari data *training* berada di kisaran 80-90%, dan pengurangan lebih jauh dari proporsi data *training*/data latih akan berdampak negatif pada kinerja model secara keseluruhan.

IV. KESIMPULAN

Berdasarkan tujuan penelitian yang difokuskan pada analisis sentimen komentar YouTube menggunakan leksikon TextBlob dan algoritma Support Vector Machine (SVM) dengan variasi rasio data uji sebesar 10%, 20%, dan 30%, dapat disimpulkan bahwa pendekatan kombinasi antara metode leksikon dan klasifikasi machine learning ini mampu digunakan untuk mengkategorikan sentimen komentar secara otomatis dan terarah. Penggunaan TextBlob sebagai alat pelabelan sentimen memberikan kemudahan dalam menilai polaritas komentar, sementara algoritma SVM menunjukkan kemampuan yang baik dalam mengklasifikasikan data yang telah dibobotkan. Variasi rasio data uji yang diterapkan turut memengaruhi akurasi model, sehingga proporsi antara data latih dan data uji terbukti menjadi faktor penting dalam performa klasifikasi. Penelitian ini menunjukkan bahwa dengan ruang lingkup yang spesifik, seperti pada satu video YouTube tentang ulasan produk, analisis sentimen dapat memberikan wawasan yang bermakna bagi produsen untuk memahami persepsi konsumen, serta mendukung pengambilan keputusan yang berbasis data.

DAFTAR PUSTAKA

[1] S. Thomas, Yuliana, dan Noviyanti. P, "Study Analisis Metode Analisis Sentimen pada YouTube," *J. Inf. Technol.*, vol. 1, no. 1, hal. 1–7, 2021, doi: 10.46229/jifotech.v1i1.201.

[2] A. Ramadhanu, R. Ayu Mahessya, M. Raihan Zaky, M. Isra, S. Informasi, dan U. Putra Indonesia YPTK Padang, "Penerapan Teknologi Machine Learning Dengan Metode Vader Pada Aplikasi Sentimen Tamu Di Hotel Dymens," *JOISIE J. Inf. Syst. Informatics Eng.*, vol. 7, no. 1, hal. 165–173, 2023.

[3] Styawati, Andi Nurkholis, Zaenal Abidin, dan Heni Sulistiani, "Optimasi Parameter Support Vector Machine Berbasis Algoritma Firefly Pada Data Opini Film," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 5, no. 5, hal. 904–910, 2021, doi: 10.29207/resti.v5i5.3380.

[4] A. Setiawan, *Analisis Sentimen Masyarakat di Twitter terhadap Kejadian Bom Bunuh Diri Polsek Astana Anyar Menggunakan Algoritma SVM dengan Leksikon Vader dan Inset*. 2024.

[5] L. Kurniasari, A. E. Cahyono, and H. L. H. Putri, "Analisis Sentimen Masyarakat terhadap Komentar Film 'Tilik' pada Media Sosial Twitter Menggunakan Metode Support Vector Machine (SVM)," *Jurnal Sisfotek Global*, vol. 11, no. 1, pp. 100–106, 2021.

[6] Y. W. H. Harahap and T. R. Widyawati, "Analisis Sentimen Ulasan Pengguna Aplikasi Mobile JKN Menggunakan TextBlob dan Naïve Bayes Classifier," *Jurnal Ilmiah Rekayasa dan Manajemen Sistem Informasi*, vol. 8, no. 1, pp. 23–32, 2022.

[7] K. A. Putri, A. T. Wibowo, and I. Mukhlash, "Analisis Sentimen Komentar Film Bioskop Menggunakan Naive Bayes Classifier dan Lexicon Based," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, vol. 4, no. 10, pp. 3855–3862, 2020.

[8] H. Suhartono, R. M. N. Rachman, and E. M. S., "Analisis Sentimen Komentar Pengguna Youtube dengan Metode Naive Bayes dan Lexicon Based Approach," *Jurnal Komputer Terapan*, vol. 8, no. 2, pp. 91–101, 2022.

[9] K. A. Putri, A. T. Wibowo, and I. Mukhlash, "Analisis Sentimen Komentar Film Bioskop Menggunakan Naive Bayes Classifier dan Lexicon Based," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, vol. 4, no. 10, pp. 3855–3862, 2020.

[10] F. Fardiansyah and M. N. K. Putra, "Analisis Sentimen Tingkat Kepuasan Pelanggan Indihome Menggunakan Algoritma Support Vector Machine (SVM)," *Jurnal Ilmiah Informatika Global*, vol. 12, no. 2, pp. 143–150, 2021.