Sentiment Analysis of YouTube Comments on the East Java Grant Case Using the Support Vector Machine (SVM) Algorithm. [Analisis Sentimen Komentar Youtube Terhadap Kasus Dana Hibah Jawa Timur Menggunakan Algoritma Support Vector Machine (SVM)]

Tirta Arya Bimantoro[1)], Yunianita Rahmawati[2], Yulian Findawati[3], M. Alfan Rosid[4]

[1)]*Program Studi Informatika, Universitas Muhammadiyah Sidoarjo, Indonesia*
[2)] *Program Studi Informatika, Universitas Muhammadiyah Sidoarjo, Indonesia*
[3)] *Program Studi Informatika, Universitas Muhammadiyah Sidoarjo, Indonesia*
[4)] *Program Studi Informatika, Universitas Muhammadiyah Sidoarjo, Indonesia*
*Email Penulis Korespondensi : yunianita@umsida.ac.id

Abstract. *The phenomenon of corruption in Indonesia, particularly the alleged misuse of grant funds in East Java Province, has sparked various public responses expressed through comments on the YouTube platform. This study aims to classify and analyze public sentiment regarding the issue by applying the Support Vector Machine (SVM) algorithm. The dataset consists of 1,039 relevant comments collected from several YouTube videos related to the case. These comments underwent a series of text preprocessing stages, including case folding, tokenizing, text cleaning, stopwords removal, stemming, and transformation using Term Frequency-Inverse Document Frequency (TF-IDF). Subsequently, the data were manually labeled into three sentiment categories: positive, negative, and neutral. The dataset was then split into training and testing sets with an 80:20 ratio and used to train and evaluate the classification model. Model evaluation was performed using accuracy, precision, recall, and F1-score metrics. The results indicate that the SVM algorithm performs well in classifying public sentiment, making it a potential tool for automatically identifying public perceptions of emerging social and political issues*

Keywords - *Corruption, YouTube, Sentiment Analysis, SVM, TF-ID*

Abstrak. *Fenomena korupsi di Indonesia, khususnya kasus dugaan penyalahgunaan dana hibah di Provinsi Jawa Timur, telah memicu berbagai tanggapan masyarakat yang banyak disampaikan melalui komentar di platform YouTube. Penelitian ini bertujuan untuk mengelompokkan dan menganalisis sentimen publik terhadap isu tersebut dengan menerapkan algoritma Support Vector Machine (SVM). Data yang digunakan berjumlah 1.039 komentar relevan yang dikumpulkan dari beberapa video YouTube terkait kasus. Komentar-komentar tersebut diproses melalui tahapan pra-pemrosesan teks, yang meliputi case folding, tokenizing, pembersihan teks, penghapusan kata tidak penting (stopwords), stemming, serta transformasi menggunakan metode Term Frequency-Inverse Document Frequency (TF-IDF). Selanjutnya, data diberi label secara manual menjadi tiga kategori sentimen: positif, negatif, dan netral. Data kemudian dibagi menjadi data latih dan data uji dengan rasio 80:20, dan digunakan untuk melatih serta menguji model klasifikasi. Evaluasi dilakukan menggunakan metrik akurasi, presisi, recall, dan f1-score. Hasil penelitian menunjukkan bahwa algoritma SVM memiliki performa yang baik dalam mengklasifikasikan sentimen publik, sehingga berpotensi digunakan sebagai alat bantu dalam mengidentifikasi persepsi masyarakat secara otomatis terhadap isu sosial dan politik yang berkembang.*

Kata Kunci - *Korupsi, YouTube, Analisis Sentimen, SVM, TF-IDF*

I. PENDAHULUAN

Korupsi merupakan permasalahan serius yang mengakar dalam sistem pemerintahan Indonesia. Meski berbagai kebijakan dan lembaga antikorupsi telah dibentuk, praktik korupsi tetap marak terjadi dari tingkat pusat hingga daerah. Salah satu kasus yang mencuat ke publik adalah dugaan penyalahgunaan dana hibah oleh oknum di Pemerintah Provinsi Jawa Timur. Kasus ini mendapat sorotan luas dari masyarakat yang mengekspresikan pendapatnya lewat berbagai platform media sosial, seperti YouTube[1].

YouTube kini tidak hanya sekedar menjadi media hiburan, melainkan juga menjadi sarana masyarakat menyampaikan opini terhadap isu sosial dan politik secara terbuka. Fenomena ini dapat dimanfaatkan melalui analisis sentimen untuk menangkap persepsi publik secara sistematis. Analisis sentimen adalah teknik dalam Natural Language Processing (NLP) yang bertujuan untuk mengidentifikasi polaritas suatu teks, apakah positif, negatif, atau netral. Dalam proses klasifikasi teks, algoritma Support Vector Machine (SVM) dikenal sebagai salah satu metode yang paling efektif dan akurat[2].

Sejumlah penelitian terdahulu menunjukkan keberhasilan penggunaan SVM dalam menganalisis komentar terhadap calon presiden Prabowo Subianto di YouTube menggunakan metode SVM, dan memperoleh akurasi sebesar

85%[1]. Desti Mualfah menggunakan SVM untuk mengklasifikasikan komentar tentang deportasi Ustadz Abdul Somad dan menunjukkan performa klasifikasi yang kuat[3]. Putri dan Cahyono menggabungkan metode penggunaan algoritma SVM dan Naïve Bayes dalam mengkaji sentimen publik terhadap kualitas pelayanan publik pemerintah DKI Jakarta, dan menemukan bahwa SVM mampu menangkap persepsi publik secara efektif[4].

Sementara itu menganalisis sentimen terhadap kasus korupsi PT. Timah menggunakan SVM dan teknik SMOTE digunakan guna menangani distribusi data yang tidak seimbang. Hasil penelitian ini membuktikan bahwa algoritma SVM tetap menunjukkan performa yang optimal meskipun data tidak seimbang[5]. Fremmuzar dan Baita melakukan pengujian terhadap berbagai jenis Penerapan kernel SVM untuk mengevaluasi sentimen pengguna terhadap layanan Telkomsel di platform Twitter dan menemukan bahwa pemilihan kernel mempengaruhi performa model[6].

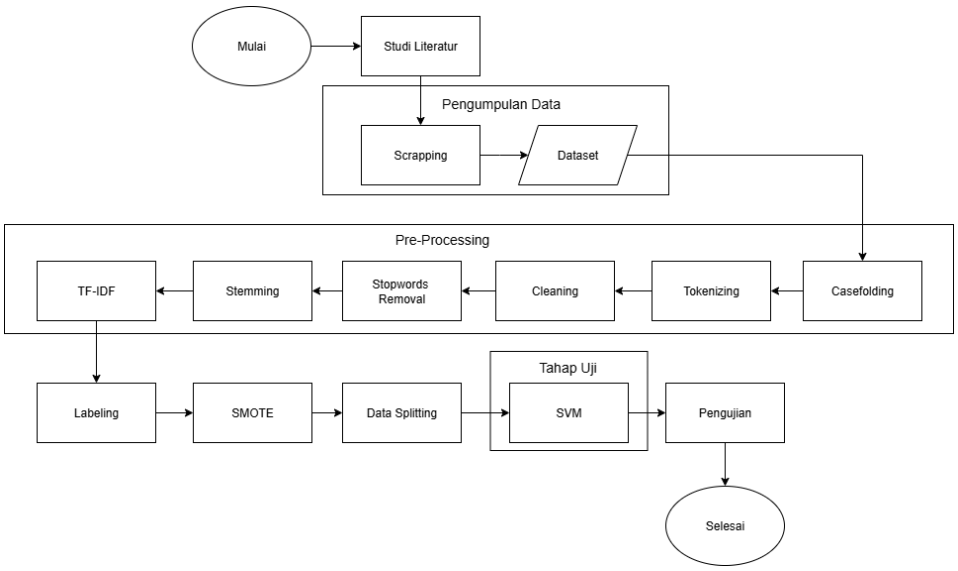
Sugandi et al. menggunakan algoritma Naïve Bayes untuk mengklasifikasikan komentar tentang kebijakan BPJS. Meskipun algoritma tersebut cukup efektif, penelitian ini tidak membahas analisis terhadap isu korupsi[7]. Hal serupa dilakukan oleh Algoritme et al. yang menganalisis komentar pada saluran beauty vlogger menggunakan SVM, menunjukkan relevansi metode ini untuk berbagai konteks sosial[2].

Kesenjangan penelitian terletak pada minimnya studi yang berfokus pada analisis sentimen komentar publik di platform YouTube terkait kasus korupsi di tingkat provinsi di Indonesia, khususnya Jawa Timur. Padahal, isu ini penting karena dapat mencerminkan persepsi nyata masyarakat terhadap praktik korupsi di daerah. Oleh karena itu, penelitian ini bertujuan untuk mengisi gap tersebut dengan menganalisis sentimen komentar YouTube terkait kasus dana hibah Jawa Timur menggunakan algoritma SVM[8].

II. METODE

2.1 Rancangan Penelitian

Desain studi adalah langkah yang akan diambil untuk menentukan penelitian dari rancangan awal hingga pengujian. Algoritma **Support Vector Machine (SVM)** digunakan untuk mengelompokkan sentimen masyarakat terhadap kasus dugaan penyalahgunaan dana hibah di Provinsi Jawa Timur berdasarkan komentar di YouTube. Dalam penelitian ini, algoritma SVM digunakan dengan kernel Radial Basis Function (RBF) karena kernel ini efektif dalam menangani data teks berdimensi tinggi. Selain itu, kernel linear juga diuji sebagai pembanding. Tahap dalam penelitian ini mencakup studi literatur, proses pengumpulan data, pra-pemrosesan teks, pelabelan manual, penyeimbangan data menggunakan **SMOTE**, pembagian dataset, pelatihan model, serta pengujian performa klasifikasi[9].



2.1.1 Kajian Pustaka

Tahap kajian pustaka disusun untuk menunjang referensi ilmiah dari beragam sumber, termasuk buku dan artikel ilmiah, dan juga media video, guna mendukung kelengkapan dan kekuatan landasan penelitian. Tahapan ini bertujuan untuk memberikan dasar teori serta pemahaman mendalam mengenai analisis sentimen, sehingga penelitian memiliki arah yang jelas dan didukung oleh kajian ilmiah yang relevan.

2.1.2 Pengumpulan Data

Sumber Dalam proses pengambilan data, komentar terkait kasus korupsi dana hibah Jawa Timur dihimpun dari platform YouTube. Pengumpulan dilakukan melalui penarikan data memakai YouTube API versi 3 melalui akun Google Console. Komentar dikumpulkan dari sejumlah video berita yang relevan dengan topik penelitian, data yang diambil dari tahun 2022 – 2025. Setelah proses scraping selesai, seluruh data disimpan dalam format CSV untuk memudahkan tahap-tahap pengolahan data berikutnya. Secara keseluruhan, jumlah data komentar yang diperoleh mencapai kurang lebih 1049 komentar. Dataset ini kemudian digunakan dalam tahap pre-processing, labelling, dan klasifikasi emosi publik menggunakan metode SVM.

Tabel 1. Pengumpulan Data

Video URL	Komentar
https://www.youtube.com/watch?v=UqMGrOcC0QM&t=73s	Jatim cuma 9M saja kapan giliran jabar 45M cum...
https://www.youtube.com/watch?v=UqMGrOcC0QM&t=73s	dftr hey kota ngawi dn
https://www.youtube.com/watch?v=UqMGrOcC0QM&t=73s	LANGGANAN KORUPSI
https://www.youtube.com/watch?v=UqMGrOcC0QM&t=73s	Jawa hama : korupsi😂
https://www.youtube.com/watch?v=UqMGrOcC0QM&t=73s	HALAH! Apa kabar MEGA KORUPSI PERTAMINA? Ya...
https://www.youtube.com/watch?v=UqMGrOcC0QM&t=73s	KPK ITU SALAH ALAMAT. GUBERNURE SING GEBLEK ITU...
https://www.youtube.com/watch?v=UqMGrOcC0QM&t=73s	Jatim gae bancaan koruptor😂😂😂 piye bu gubernurnya
https://www.youtube.com/watch?v=UqMGrOcC0QM&t=73s	Knp Dana Hibah ke Jatim banyak bermasalah ... ...
https://www.youtube.com/watch?v=UqMGrOcC0QM&t=73s	Yg dulu merasa aman kini duduk pun tak tenang ...
https://www.youtube.com/watch?v=UqMGrOcC0QM&t=73s	Berantas semua penghianat negeri ini monyet2 d...

2.1.3 Pre-Processing

Sebelum data digunakan untuk proses pelabelan dan klasifikasi sentimen, dilakukan tahap pra-pemrosesan guna membersihkan dan mempersiapkan teks komentar. Tahapan ini bertujuan untuk mengubah data mentah menjadi format yang lebih bersih, konsisten, dan terstruktur sehingga dapat dianalisis secara optimal oleh algoritma machine learning seperti Support Vector Machine (SVM). Proses ini sangat penting mengingat karakteristik komentar di media sosial yang sering kali tidak terstandar, penuh dengan variasi bahasa, simbol, dan elemen non-teks lainnya yang dapat mengganggu akurasi analisis. Adapun tahapan yang dilakukan adalah sebagai berikut sebagai berikut.

a. Casefolding

Semua huruf dalam komentar diubah menjadi huruf kecil (*lowercase*) guna menghindari perbedaan penulisan pada kata yang sama namun ditulis dengan kapitalisasi berbeda.

Tabel 2. Casefolding

Komentar	Casefolding
Jatim cuma 9M saja kapan giliran jabar 45M cum...	jatim cuma 9m saja kapan giliran jabar 45m cum...
dftr hey kota ngawi dn	dftr hey kota ngawi dn
LANGGANAN KORUPSI	langganan korupsi
Jawa hama : korupsi😂	jawa hama : korupsi😂
HALAH! Apa kabar MEGA KORUPSI PERTAMINA? Ya...	halah! apa kabar mega korupsi pertamina? ya...
KPK ITU SALAH ALAMAT. GUBERNURE SING GEBLEK ITU...	kpk itu salah alamat. gubernure sing geblek itu...
Jatim gae bancaan koruptor😂😂😂 piye bu gubernurnya	jatim gae bancaan koruptor😂😂😂 piye bu gubernurnya

Knp Dana Hibah ke Jatim banyak bermasalah	knp dana hibah ke jatim banyak bermasalah ... ...
Yg dulu merasa aman kini duduk pun tak tenang ...	yg dulu merasa aman kini duduk pun tak tenang ...
Berantas semua penghianat negeri ini monyet2 d...	berantas semua penghianat negeri ini monyet2 d...

b. Tokenizing

Proses ini memisahkan kalimat atau paragraf menjadi unit kata tunggal (*tokens*). Tokenizing penting untuk memecah informasi menjadi elemen-elemen dasar yang akan dianalisis.

Tabel 3. Tokenizing

Casefolding	Tokenizing
jatim cuma 9m saja kapan giliran jabar 45m cum...	[jatim, cuma, 9m, saja, kapan, giliran, jabar,...
dftr hey kota ngawi dn	[dftr, hey, kota, ngawi, dn]
langganan korupsi	[*, langganan, korupsi, *]
jawa hama : korupsi😂	[jawa, hama, :, korupsi😂]
halah! apa kabar mega korupsi pertamina? ya...	[halah, !, apa, kabar, mega, korupsi, pertamin...
kpk itu salah alamat. gubernure sing geblek itu...	[kpk, itu, salah, alamat, ., gubernure, sing, ...
jatim gae bancaan koruptor😂😂😂 piye bu gubernurnya	[jatim, gae, bancaan, koruptor😂😂😂, piye, bu, g...
knp dana hibah ke jatim banyak bermasalah	[knp, dana, hibah, ke, jatim, banyak, bermasal...
yg dulu merasa aman kini duduk pun tak tenang ...	[yg, dulu, merasa, aman, kini, duduk, pun, tak...
berantas semua penghianat negeri ini monyet2 d...	[berantas, semua, penghianat, negeri, ini, mon...

c. Cleaning

Tahap ini melibatkan penghapusan karakter atau elemen yang tidak relevan dalam teks seperti tanda baca, angka, URL, mention (@username), hashtag, emotikon, serta karakter non-alfabet lainnya. Cleaning penting agar hanya informasi inti yang diproses lebih lanjut.

Tabel 4. Cleaning

Tokenizing	Cleaning
[jatim, cuma, 9m, saja, kapan, giliran, jabar,...	[jatim, cuma, m, saja, kapan, giliran, jabar, ...
[dftr, hey, kota, ngawi, dn]	[dftr, hey, kota, ngawi, dn]
[*, langganan, korupsi, *]	[langganan, korupsi]
[jawa, hama, :, korupsi😂]	[jawa, hama, korupsi]
[halah, !, apa, kabar, mega, korupsi, pertamin...	[halah, apa, kabar, mega, korupsi, pertamina, ...
[kpk, itu, salah, alamat, ., gubernure, sing, ...	[kpk, itu, salah, alamat, gubernure, sing, geb..
[jatim, gae, bancaan, koruptor😂😂😂, piye, bu, g...	[jatim, gae, bancaan, koruptor, piye, bu, gube...
[knp, dana, hibah, ke, jatim, banyak, bermasal...	[knp, dana, hibah, ke, jatim, banyak, bermasal...
[yg, dulu, merasa, aman, kini, duduk, pun, tak...	[yg, dulu, merasa, aman, kini, duduk, pun, tak...
[berantas, semua, penghianat, negeri, ini, mon...	[berantas, semua, penghianat, negeri, ini, mon...

d. Stopword Removal

Stopword Removal adalah tahap pembersihan teks dengan cara menghilangkan kata-kata yang bersifat umum dan tidak memiliki pengaruh signifikan terhadap makna atau konteks utama dalam sebuah teks seperti “yang”, “di”, “dengan”, “pada”, “untuk”. Stopword dihapus agar model hanya fokus pada kata-kata bermakna utama. Daftar stopword bahasa Indonesia dapat diambil dari pustaka NLP seperti Sastrawi atau NLTK.

Tabel 5. Stopword Removal

Cleaning	Stopword Removal
[jatim, cuma, m, saja, kapan, giliran, jabar, ...	[jatim, m, giliran, jabar, m, yayasan, mencapa...
[dftr, hey, kota, ngawi, dn]	[dftr, hey, kota, ngawi, dn]
[langganan, korupsi]	[langganan, korupsi]
[jawa, hama, korupsi]	[jawa, hama, korupsi]
[halah, apa, kabar, mega, korupsi, pertamina, ...	[halah, kabar, mega, korupsi, pertamina, dipus...
[kpk, itu, salah, alamat, gubernure, sing, geb..	[kpk, salah, alamat, gubernure, sing, geblek, ...
[jatim, gae, bancaan, koruptor, piye, bu, gube...	[jatim, gae, bancaan, koruptor, piye, bu, gube...
[knp, dana, hibah, ke, jatim, banyak, bermasal...	[knp, dana, hibah, jatim, bermasalah, salah]
[yg, dulu, merasa, aman, kini, duduk, pun, tak...	[yg, aman, duduk, tenang, yo, rasakno, kusus, ...
[berantas, semua, penghianat, negeri, ini, mon...	[berantas, penghianat, negeri, monyet, dableg,...

e. Stemming

Proses stemming mengubah kata turunan menjadi bentuk dasarnya. Misalnya, kata “menyalahgunakan”, “penyalahgunaan”, dan “disalahgunakan” semuanya akan distem menjadi “salahguna”. Teknik ini membantu mengurangi dimensi fitur dan menyatukan kata dengan akar yang sama. Proses ini dilakukan menggunakan stemmer bahasa Indonesia seperti Sastrawi.

Tabel 6. Stemming

Stopword Removal	Stemming
[jatim, m, giliran, jabar, m, yayasan, mencapa...	[jatim, m, gilir, jabar, m, yayasan, capai, tr...
[dftr, hey, kota, ngawi, dn]	[dftr, hey, kota, ngawi, dn]
[langganan, korupsi]	[langgan, korupsi]
[jawa, hama, korupsi]	[jawa, hama, korupsi]
[halah, kabar, mega, korupsi, pertamina, dipus...	[halah, kabar, mega, korupsi, pertamina, dipus...
[kpk, salah, alamat, gubernure, sing, geblek, ...	[kpk, salah, alamat, gubernure, sing, geblek, ...
[jatim, gae, bancaan, koruptor, piye, bu, gube...	[jatim, gae, bancaan, koruptor, piye, bu, gube...
[knp, dana, hibah, jatim, bermasalah, salah]	[knp, dana, hibah, jatim, masalah, salah]
[yg, aman, duduk, tenang, yo, rasakno, kusus, ...	[yg, aman, duduk, tenang, yo, rasakno, sus, ja...
[berantas, penghianat, negeri, monyet, dableg,...	[berantas, penghianat, negeri, monyet, dableg,...

f. TF-IDF

Setelah teks dibersihkan dan distem, langkah selanjutnya adalah mengubah kumpulan kata diubah menjadi vektor numerik. Pendekatan Term Frequency-Inverse Document Frequency (TF-IDF)

dimanfaatkan untuk menetapkan bobot tinggi pada kata-kata yang sering muncul dalam satu komentar, tetapi jarang ditemukan di seluruh kumpulan data. Strategi ini bertujuan untuk mengidentifikasi kata-kata yang paling mewakili isi masing-masing dokumen. Nilai TF-IDF inilah yang kemudian digunakan sebagai fitur dalam model klasifikasi.

hasil TF-IDF (10 komentar pertama):

	aamiin	aamin	abah	abal	abangkuh	abangkuhh	abangkuuu	abis	abiz	abs	...	yme	yng	yo	yq	yra	yth	yuk	zet	zubair	zulhas
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0
5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0
6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0
7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0
8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.309361	0.0	0.0	0.0	0.0	0.0	0.0	0.0
9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0

10 rows × 2807 columns

Hasil TF-IDF disimpan sebagai 'hasil_tfidf.csv'

Gambar 1. Hasil TF-IDF

2.1.4 Labelling

Setelah melalui pra-pemrosesan, setiap komentar dikelompokkan secara manual ke dalam tiga kategori sentimen utama, yakni positif, negatif, dan netral, berdasarkan isi dan konteks kalimat. Proses ini bertujuan membentuk dataset yang akurat dan seimbang untuk pelatihan dan pengujian model. Teknik pelabelan ini juga diterapkan oleh Abdulloh & Pambudi dalam studi mereka[10].

Tabel 7. Labelling

Komentar	Labelling
Jatim cuma 9M saja kapan giliran jabar 45M cum...	-1
dftr hey kota ngawi dn	-1
LANGGANAN KORUPSI	-1
Jawa hama : korupsi😂	-1
HALAH! Apa kabar MEGA KORUPSI PERTAMINA? Ya...	-1
KPK ITU SALAH ALAMAT. GUBERNURE SING GEBLEK ITU...	-1
Jatim gae bancaan koruptor😂😂😂 piye bu gubernurnya	-1
Knp Dana Hibah ke Jatim banyak bermasalah	-1
Yg dulu merasa aman kini duduk pun tak tenang ...	-1
Berantas semua penghianat negeri ini monyet2 d...	-1

2.1.5 Smote

Data awal menunjukkan ketimpangan distribusi kelas, di mana sentimen negatif jumlahnya lebih banyak dibanding kelas lainnya. Hal ini dapat mengakibatkan model lebih condong terhadap kelas mayoritas. Untuk mengatasi ketimpangan ini, digunakan teknik SMOTE (Synthetic Minority Over-sampling Technique), sebuah metode untuk memperbanyak data sintetis dari kelas minoritas berdasarkan jarak antar tetangga terdekat.

Rumus dasar SMOTE :

$$x_{\text{baru}} = x_i + \delta \times (x_{zi} - x_i)$$

Keterangan :
Xi : titik data minoritas
Xzi : tetangga terdekat dari Xi
δ : angka acak antara 0 dan 1
Teknik ini terbukti efektif dalam meningkatkan kinerja model klasifikasi pada data tidak seimbang, seperti dijelaskan dalam studi oleh Asrianto & Herwinanda [11].

2.1.6 Data Splitting

Setelah proses SMOTE, data kemudian dibagi melalui tahap data splitting :

- 80% digunakan untuk pelatihan model (*training data*).
- 20% digunakan untuk menguji performa model (*testing data*).

Pembagian ini dilakukan setelah proses balancing agar data pelatihan memiliki distribusi sentimen yang proporsional, sebagaimana dilakukan pula oleh Aprilia & Isnain[12]

2.1.7 Evaluasi Model

Model klasifikasi dievaluasi menggunakan empat metrik utama untuk mengukur performa secara menyeluruh, yaitu: akurasi, presisi, recall, dan f1-score. Kombinasi metrik tersebut digunakan untuk memberikan penilaian yang menyeluruh terhadap performa model, terutama ketika menangani data yang tidak seimbang atau mengandung noise[13]. Berdasarkan hasil pengujian, kernel RBF menunjukkan performa klasifikasi yang lebih optimal dibandingkan kernel linear, sehingga digunakan sebagai konfigurasi utama dalam penelitian ini.

Rumus Evaluasi :

- **Akurasi :**

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

- **Presisi :**

$$\text{Precision} = \frac{TP}{TP + FP}$$

- **Recall :**

$$\text{Recall} = \frac{TP}{TP + FN}$$

- **F1-Score :**

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Keterangan :

- *TP* = True Positive
- *TN* = True Negative
- *FP* = False Positive
- *FN* = False Negative

III. HASIL DAN PEMBAHASAN

Proses klasifikasi sentimen terhadap komentar masyarakat dalam kasus dugaan penyalahgunaan dana hibah di Provinsi Jawa Timur dilakukan dengan menggunakan algoritma Support Vector Machine (SVM). Tahapan analisis dimulai dari eksplorasi distribusi awal sentimen, yang menunjukkan ketidakseimbangan signifikan, di mana sebagian besar komentar bersifat negatif. Untuk mengatasi masalah ini, diterapkan teknik Synthetic Minority Over-sampling Technique (SMOTE) guna menghasilkan data sintetik pada kelas minoritas agar distribusi antar kelas menjadi seimbang. Selanjutnya, Setelah melalui proses pra-pemrosesan, data dibagi menjadi dua subset, yakni data untuk pelatihan dan data untuk pengujian. Model SVM dilatih menggunakan data pelatihan tersebut, lalu kinerjanya dievaluasi menggunakan sejumlah metrik, termasuk akurasi, presisi, recall, dan f1-score, untuk mengukur tingkat ketepatan dalam mengklasifikasikan komentar.

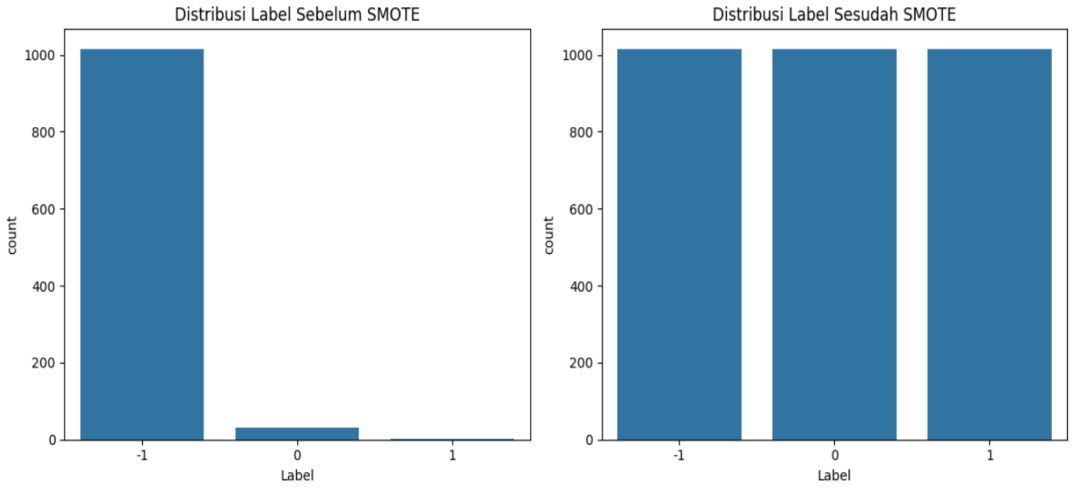
Hasil evaluasi model SVM menunjukkan kinerja klasifikasi yang sangat baik pada seluruh kategori sentimen, yaitu positif, netral, dan negatif. Untuk menggambarkan secara rinci jumlah prediksi yang tepat dan keliru pada setiap kelas, digunakan confusion matrix. Untuk memperkuat pemahaman terhadap isi komentar, dilakukan juga visualisasi dengan wordcloud yang menampilkan kata-kata dominan dalam setiap kategori sentimen. Visualisasi ini memperlihatkan bahwa komentar negatif didominasi kata seperti "korupsi" dan "uang", sedangkan komentar netral cenderung informatif, dan komentar positif memuat unsur dukungan dan harapan. Secara keseluruhan, penerapan model SVM yang dipadukan dengan teknik penyeimbangan data dan analisis visual teks berhasil mengidentifikasi opini publik secara otomatis melalui platform digital seperti YouTube[14].

3.1 Distribusi Awal dan Penyeimbangan Data dengan SMOTE

Distribusi awal data sentimen menunjukkan adanya ketimpangan kelas yang sangat ekstrem, di mana sebagian besar komentar tergolong negatif. Dari total 1.049 komentar yang dikumpulkan dari berbagai video YouTube yang relevan, hanya 3 komentar tergolong positif, 30 tergolong netral, dan sisanya sebanyak 1.016 komentar masuk kategori negatif. Ketidakseimbangan semacam ini merupakan masalah umum dalam klasifikasi teks berbasis opini publik, khususnya pada isu yang sensitif atau kontroversial seperti korupsi. Hal ini berpotensi menyebabkan model klasifikasi mengalami bias terhadap kelas mayoritas dan mengabaikan pola pada kelas minoritas.

Untuk mengatasi masalah tersebut, digunakan teknik SMOTE (Synthetic Minority Over-sampling Technique) yang merupakan pendekatan populer dalam penyeimbangan data. SMOTE bekerja dengan menciptakan contoh data sintetis untuk kelas minoritas dengan cara interpolasi dari tetangga terdekat dalam ruang fitur. Teknik ini lebih efektif daripada duplikasi data, karena menghasilkan sampel baru yang menyerupai distribusi kelas minoritas tanpa mengulang data yang sama[15].

Setelah penerapan SMOTE, jumlah data pada masing-masing kelas menjadi seimbang, yaitu masing-masing 1.016 data untuk sentimen positif, netral, dan negatif. Distribusi yang seimbang ini sangat penting agar algoritma SVM dapat belajar secara merata dari setiap kelas, meningkatkan kemampuan generalisasi model, dan meminimalkan bias klasifikasi. Visualisasi perbandingan distribusi sebelum dan sesudah SMOTE ditampilkan pada Gambar 2



Gambar 2. Distribusi Label Sebelum dan Sesudah SMOTE

3.2 Evaluasi Performa Model SVM

Setelah data diseimbangkan, dataset dibagi menjadi dua bagian: 80% digunakan untuk pelatihan dan 20% untuk pengujian. Algoritma SVM dilatih menggunakan data pelatihan dengan fitur teks yang telah diolah melalui metode TF-IDF. Evaluasi model dilakukan menggunakan empat metrik utama akurasi, presisi, recall, dan f1-score karena metrik-metrik ini mampu memberikan penilaian komprehensif terhadap kinerja model, khususnya dalam klasifikasi teks dengan banyak kategori.

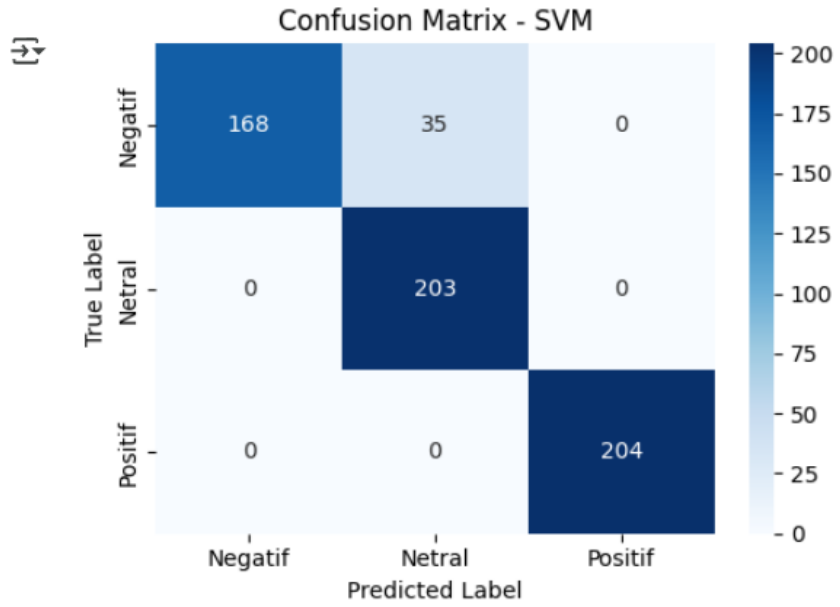
Tabel 8. Hasil Evaluasi Model SVM

Metrik	Negatif (-1)	Netral (0)	Positif (1)	Rata-rata (Macro avg)
Precision	1.00	0.85	1.00	0.95
Recall	0.83	1.00	1.00	0.94
F1-Score	0.91	0.92	1.00	0.94
Support	203	203	204	610
Akurasi	–	–	–	**0.94 / 94.26%**

Berdasarkan tabel di atas, dapat disimpulkan bahwa model SVM memiliki performa klasifikasi yang sangat baik untuk semua kategori sentimen. Meskipun sebelumnya data bersifat tidak seimbang, penerapan SMOTE memungkinkan model mempelajari pola dari ketiga kelas secara proporsional. Hal ini tercermin dari nilai F1-Score yang tinggi dan merata di semua label. Hasil ini menunjukkan bahwa model tidak hanya mampu mengenali komentar negatif, tetapi juga sensitif terhadap komentar positif dan netral, yang sering kali diabaikan oleh model pada dataset tidak seimbang.

3.3 Visualisasi Confusion Matrix

Untuk memperoleh pemahaman yang lebih mendalam terhadap kinerja model klasifikasi, digunakan confusion matrix sebagai alat evaluasi. Matriks ini menunjukkan perbandingan antara hasil prediksi model dan label sebenarnya, sehingga dapat mengidentifikasi secara jelas jenis kesalahan yang terjadi. Berdasarkan visualisasi yang ditampilkan pada Gambar 3, mayoritas prediksi model berada di diagonal utama, yang menandakan tingkat akurasi yang tinggi karena banyak prediksi sesuai dengan label aslinya. Kesalahan prediksi seperti false positive dan false negative jumlahnya relatif kecil dan tersebar merata di luar diagonal, menunjukkan bahwa model tidak berat sebelah terhadap satu kelas tertentu. Dengan demikian, penerapan algoritma SVM yang dilengkapi dengan metode penyeimbangan data menggunakan SMOTE terbukti efektif dalam mengurangi kesalahan klasifikasi dan menghasilkan performa yang stabil di seluruh kelas.



Gambar 3. Confusion Matrix Klasifikasi Sentimen Menggunakan SVM

3.4 Visualisasi Wordcloud Sentimen

Wordcloud digunakan untuk menampilkan kata-kata yang paling sering digunakan dalam setiap kategori sentimen. Teknik ini sangat berguna untuk memahami isi dan konteks komentar tanpa harus membaca satu per satu. Setiap kategori ditampilkan dalam visualisasi terpisah agar lebih mudah dibedakan, yaitu Gambar 4 untuk komentar negatif, Gambar 5 untuk komentar netral, dan Gambar 6 untuk komentar positif.

Gambar 4. WordCloud Sentimen Negatif

Terlihat bahwa kata-kata seperti “korupsi”, “penjara”, “rakyat”, dan “uang” muncul secara dominan. Kemunculan kata-kata ini menunjukkan bahwa masyarakat memiliki persepsi yang kuat dan emosional terhadap isu korupsi, terutama dalam konteks penyalahgunaan dana hibah. Kata "korupsi" sebagai kata kunci utama mencerminkan pusat perhatian publik, sementara kata "penjara" menunjukkan adanya harapan agar pelaku dihukum secara tegas. Kata "rakyat" menandakan bahwa masyarakat merasa menjadi pihak yang dirugikan atau korban dari praktik korupsi tersebut. Sementara itu, kemunculan kata "uang" menekankan bahwa isu ini sangat sensitif karena menyangkut dana publik yang seharusnya digunakan untuk kepentingan umum.

Gambar 5. WordCloud Sentimen Netral

Menampilkan kata seperti “kasus”, “laporan”, “hibah”, dan “data”. Kata-kata ini bersifat netral, deskriptif, dan sering kali muncul dalam komentar yang hanya menyampaikan informasi atau bertanya tanpa muatan emosi

Gambar 6. WordCloud Sentimen Positif

Meskipun lebih sedikit jumlahnya, wordcloud kategori ini menunjukkan kata-kata seperti “adil”, “transparan”, “baik”, dan “terima kasih”. Ini mencerminkan harapan publik terhadap penegakan hukum yang jujur dan apresiasi terhadap pihak yang dinilai bersikap benar.

3.5 Interpretasi dan Implikasi

Secara keseluruhan, temuan penelitian ini menunjukkan bahwa metode yang menggunakan Support Vector Machine (SVM) untuk klasifikasi sentimen terbukti efektif dalam menganalisis opini publik terhadap isu sosial-politik yang sensitif. Penerapan teknik SMOTE sebagai tahap penyeimbang data juga terbukti penting dalam mengurangi bias terhadap kelas mayoritas. Model mampu mengenali dengan akurat pola-pola sentimen positif dan netral yang sebelumnya minoritas.

Implikasinya, pendekatan ini dapat diterapkan secara lebih luas untuk monitoring opini publik terhadap isu kebijakan, pemilu, skandal politik, dan lainnya. Pemerintah atau lembaga pengawas dapat menggunakan sistem semacam ini untuk mengumpulkan insight dari media sosial secara otomatis dan efisien. Selain itu, pendekatan ini juga relevan dalam penelitian sosial digital dan bidang komunikasi politik.

Penelitian ini menunjukkan bahwa penggunaan algoritma SVM yang dikombinasikan dengan teknik balancing seperti SMOTE dan visualisasi berbasis NLP (Natural Language Processing) dapat menjadi alat penting dalam era keterbukaan informasi dan partisipasi digital.

IV. KESIMPULAN

Penelitian ini menerapkan algoritma Support Vector Machine (SVM) dalam proses klasifikasi sentimen terhadap komentar masyarakat di platform YouTube terkait kasus dugaan korupsi dana hibah di Provinsi Jawa Timur. Proses dimulai dari pengumpulan data melalui YouTube API v3, dilanjutkan dengan pra-pemrosesan teks, pelabelan manual, penyeimbangan data menggunakan teknik SMOTE, hingga tahap pelatihan serta pengujian performa model klasifikasi.

Hasil pengujian menunjukkan bahwa algoritma SVM dapat mengidentifikasi sentimen dengan akurasi tinggi sebesar **94,26%**, hasil evaluasi mengindikasikan bahwa ketiga metrik utama precision, recall, dan f1-score menunjukkan performa yang konsisten dan merata di seluruh kategori sentimen, yaitu (positif, netral, dan negatif). Teknik SMOTE terbukti efektif dalam menangani ketidakseimbangan data, yang menjadi kendala umum dalam analisis sentimen berbasis media sosial. Selain itu, visualisasi data melalui confusion matrix dan wordcloud memberikan pemahaman yang lebih mendalam terhadap persepsi publik terhadap isu yang sedang berkembang.

Secara keseluruhan, pendekatan ini membuktikan bahwa kombinasi teknik pre-processing, balancing, dan pemilihan algoritma yang tepat dapat digunakan secara efektif untuk memahami opini masyarakat secara otomatis. Diharapkan bahwa penelitian ini akan membantu dalam pembangunan sistem analisis sentimen dalam

domain sosial-politik. Selain itu, akan menjadi acuan untuk penelitian terkait pemetaan persepsi publik di ruang digital.

UCAPAN TERIMA KASIH

Dengan tulus dan penuh hormat, penulis menyampaikan penghargaan yang sebesar-besarnya kepada seluruh pihak yang telah memberikan dorongan, bantuan, serta kontribusi berarti dalam setiap tahap penyusunan artikel ini. Segala bentuk perhatian dan dukungan yang diberikan menjadi fondasi penting dalam menyelesaikan artikel ini dengan baik, Yakni :

1. Bapak Dr. Hidayatulloh, M.Si, selaku Rektor Universitas Muhammadiyah Sidoarjo, atas dukungan dan arahannya selama masa studi.
2. Bapak Irwanto, ST., M.MT., Dekan Fakultas Sains dan Teknologi, yang telah memberikan motivasi serta bimbingan akademik yang berharga.
3. Ibu Ade Eviyanti, S.Kom., M.Kom., Ketua Program Studi Informatika, atas dukungan dan semangat yang diberikan sepanjang proses perkuliahan.
4. Yunianita Rahmawati., S.Kom., M.Kom., selaku dosen pembimbing, atas kesabaran dan ketelitiannya dalam membimbing penulis selama penelitian hingga penyusunan artikel ini.
5. Seluruh dosen di lingkungan Program Studi Informatika yang telah berbagi ilmu, pengalaman, dan inspirasi yang tak ternilai selama masa perkuliahan.
6. Ibu penulis, atas kasih sayang, pengorbanan, dan doa yang tiada henti. Terima kasih telah menjadi sumber kekuatan, semangat, serta menjadi "donatur utama" dalam setiap proses perjuangan ini.
7. Pasangan penulis, yang selalu ada memberi semangat, pengertian, dan dukungan emosional di setiap fase sulit. Terima kasih telah menjadi penyemangat sekaligus teman berpikir dalam proses panjang ini.
8. Semua pihak yang tidak dapat disebutkan satu per satu, yang telah memberikan bantuan, motivasi, dan dukungan dalam berbagai bentuk.

Semoga segala bentuk dukungan, kebaikan, dan bantuan yang telah diberikan mendapat balasan terbaik dari Allah SWT. Artikel ini tidak akan pernah selesai tanpa doa dan semangat dari orang-orang luar biasa di sekeliling penulis.

REFERENSI

[1] E. Tohidi, R. Perdana Herdiansyah, And E. Wahyudin, “Analisa Sentimen Komentar Video Youtube Di Channel Tvonenews Tentang Calon Presiden Prabowo Subianto Menggunakan Support Vector Machine,” 2024. [Online]. Available: Https://Www.Youtube.Com/Watch?V=Wu_Wqhg1ylk

[2] R. T. Adek, Z. Fitri, And S. Chairani Siegar, “Analisis Sentimen Komentar Pada Saluran Youtube Beauty Vlogger Berbahasa Indonesia Menggunakan Metode Support Vector Machine,” *Jurnal Algoritme*, Vol. 5, No. 2, Pp. 164–175, 2025, Doi: 10.35957/Algoritme.V5i2.9692.

[3] Desti Mualfah, “Analisis Sentimen Komentar Youtube Tvone Tentang Ustadz Abdul Somad Dideportasi Dari Singapura Menggunakan Algoritma Svm,” 2023.

[4] R. R. Putri And N. Cahyono, “Analisis Sentimen Komentar Masyarakat Terhadap Pelayanan Publik Pemerintah Dki Jakarta Dengan Algoritma Super Vector Machine Dan Naive Bayes,” 2024.

[5] F. Caroline, R. G. S. Budi, And M. E. Al Rivan, “Analisis Sentimen Masyarakat Terhadap Kasus Korupsi Pt. Timah Menggunakan Metode Support Vector Machine,” *Jurnal Ilmu Komputer Dan Informatika*, Vol. 4, No. 1, Pp. 43–50, Jul. 2024, Doi: 10.54082/Jiki.141.

[6] P. Fremmuzar And A. Baita, “Uji Kernel Svm Dalam Analisis Sentimen Terhadap Layanan Telkomsel Di Media Sosial Twitter,” *Komputika : Jurnal Sistem Komputer*, Vol. 12, No. 2, Pp. 57–66, Sep. 2023, Doi: 10.34010/Komputika.V12i2.9460.

[7] Hayati, “Analisis Sentimen Komentar Pengguna Youtube Terhadap Kebijakan Baru Badan Penyelenggara Jaminan Kesehatan Sosial Menggunakan Naïve Bayes,” 2024.

[8] A. Irma Purnamasari And I. Ali, “Analisis Sentimen Komentar Berita Detik.Com Menggunakan Algoritma Suport Vektor Machine (Svm),” 2024.

[9] N. Kenneth, “Maraknya Kasus Korupsi Di Indonesia Tahun Ke Tahun,” *Nathanael Kenneth*, Vol. 2, No. 1, 2024.

[10] F. F. Abdulloh And I. R. Pambudi, “Analisis Sentimen Pengguna Youtube Terhadap Program Vaksin Covid-19,” *Csrid (Computer Science Research And Its Development Journal)*, Vol. 13, No. 3, P. 141, Nov. 2021, Doi: 10.22303/Csrid.13.3.2021.141-148.

[11] R. Asrianto And M. Herwinanda, "Analisis Sentimen Kenaikan Harga Kebutuhan Pokok Dimedia Sosial Youtube Menggunakan Algoritma Support Vector Machine," *Jurnal Coscitech (Computer Science And Information Technology)*, Vol. 3, No. 3, Pp. 431–440, Dec. 2022, Doi: 10.37859/Coscitech.V3i3.4368.
[12] Aprilia And A. R. Isnain, "Analisis Sentimen Terhadap Media Sosial Twitter Dengan Kasus Kampanye Anti-Korupsi Di Indonesia Menggunakan Naive Bayes," *Jurnal Media Informatika Budidarma*, Vol. 8, No. 2, P. 695, Apr. 2024, Doi: 10.30865/Mib.V8i2.7582.
[13] Muhammad Hilmy Riwanto, "Analisis Sentimen Komentar Youtube Terkait Penerapan Makan Bergizi Gratis Menggunakan Model Algoritma Svm," 2025.
[14] L. Rofiqi And M. Akbar, "Analisis Sentimen Terkait Ruu Perampasan Aset Dengan Support Vector Machine," *Jekin - Jurnal Teknik Informatika*, Vol. 4, No. 3, Pp. 529–538, Aug. 2024, Doi: 10.58794/Jekin.V4i3.824.
[15] H. Hidayat, F. Santoso, And L. F. Lidimillah, "Analisis Sentimen Pengguna Youtube Tentang Rohingya Menggunakan Algoritma Svm (Support Vector Machine)," *G-Tech: Jurnal Teknologi Terapan*, Vol. 8, No. 3, Pp. 1729–1738, Jul. 2024, Doi: 10.33379/Gtech.V8i3.4497.

Conflict of Interest Statement:
The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.