

Artikel Penelitian (Teknik Informatika)

Analisis Sentimen Komentar Youtube terhadap Kasus Dugaan Korupsi Migas Menggunakan Algoritma Support Vector Machine (SVM)

Dion Bona Torang Sinaga *, Agus Sulistyanto, Hidayatul Ichwan

Teknik Informatika, Sekolah Tinggi Manajemen Informatika & Komputer Jayakarta, Jakarta, Indonesia

INFORMASI ARTIKEL

Diterima Redaksi: 13 Juni 2026
Revisi Akhir: 17 Juli 2026
Diterbitkan Online: 31 Juli 2026

KATA KUNCI

Oil and Gas Corruption
Sentiment Analysis
Support Vector Machine (SVM)
TF-IDF
Text Mining
YouTube

KORESPONDENSI (*)

E-mail: 22570011@stmik.jayakarta.ac.id

A B S T R A K

Perkembangan media sosial telah menjadikan platform YouTube sebagai salah satu sarana utama masyarakat dalam menyampaikan opini terhadap berbagai isu publik, termasuk kasus dugaan korupsi di sektor minyak dan gas (migas). Komentar yang dihasilkan pengguna YouTube mengandung informasi mengenai persepsi dan sentimen masyarakat, namun jumlahnya yang besar serta sifatnya yang tidak terstruktur menyebabkan analisis secara manual menjadi kurang efektif. Oleh karena itu, diperlukan metode analisis sentimen yang mampu mengolah dan mengklasifikasikan opini masyarakat secara otomatis dan akurat. Penelitian ini bertujuan untuk menganalisis sentimen komentar YouTube terhadap kasus dugaan korupsi migas menggunakan algoritma Support Vector Machine (SVM). Penelitian ini menggunakan pendekatan kuantitatif dengan tahapan pengumpulan data komentar YouTube, preprocessing data yang meliputi cleaning, case folding, tokenizing, stopword removal, dan stemming, kemudian ekstraksi fitur menggunakan Term Frequency-Inverse Document Frequency (TF-IDF). Selanjutnya, proses klasifikasi dilakukan menggunakan algoritma Support Vector Machine (SVM) dengan empat jenis kernel, yaitu linear, sigmoid, Radial Basis Function (RBF), dan polynomial. Evaluasi model dilakukan menggunakan confusion matrix dengan metrik accuracy, precision, recall, dan F1-score. Hasil penelitian menunjukkan bahwa algoritma SVM mampu mengklasifikasikan sentimen komentar YouTube ke dalam kategori positif, negatif, dan netral dengan baik. Berdasarkan hasil pengujian, kernel linear menghasilkan performa terbaik dengan nilai akurasi sebesar 74,95% dan weighted F1-score sebesar 0,75. Sementara itu, kernel sigmoid memperoleh akurasi sebesar 72,40%, kernel RBF sebesar 71,23%, dan kernel polynomial sebesar 56,45%. Hasil tersebut menunjukkan bahwa karakteristik data sentimen komentar YouTube pada kasus dugaan korupsi migas cenderung dapat dipisahkan secara linear sehingga kernel linear menjadi pilihan yang paling optimal. Penelitian ini membuktikan bahwa kombinasi TF-IDF dan SVM dapat digunakan secara efektif untuk menganalisis opini masyarakat terhadap isu publik yang berkembang di media sosial.

PENDAHULUAN

Perkembangan teknologi informasi dan komunikasi dalam beberapa tahun terakhir mengalami peningkatan yang sangat pesat, terutama dalam pemanfaatan internet dan media sosial sebagai sarana utama pertukaran informasi. Transformasi digital tersebut telah mengubah pola komunikasi masyarakat dari yang sebelumnya bersifat konvensional menjadi lebih terbuka, interaktif, dan berlangsung secara real-time. Media sosial tidak hanya berfungsi sebagai media penyebaran informasi, tetapi juga menjadi ruang publik digital yang memungkinkan masyarakat menyampaikan opini, aspirasi, kritik, maupun tanggapan terhadap berbagai isu yang berkembang di tengah kehidupan sosial. Salah satu platform media sosial yang memiliki pengaruh besar dalam penyebaran informasi adalah YouTube. Selain berfungsi sebagai media berbagi konten audiovisual, YouTube menyediakan fitur komentar yang memungkinkan pengguna memberikan respons secara

langsung terhadap video yang ditonton. Kolom komentar tersebut menjadi sumber data yang kaya akan informasi karena mencerminkan persepsi, opini, serta emosi masyarakat terhadap suatu topik tertentu. Berbagai penelitian menunjukkan bahwa komentar YouTube dapat dimanfaatkan sebagai sumber data yang efektif untuk memahami opini publik secara luas dan dinamis melalui pendekatan analisis sentimen [1][2].

Meskipun demikian, volume komentar yang sangat besar dan terus bertambah setiap waktu menimbulkan tantangan dalam proses pengolahan dan analisis data. Komentar pada media sosial umumnya berbentuk data teks tidak terstruktur yang mengandung bahasa informal, singkatan, kesalahan penulisan, serta variasi kosakata yang tinggi. Kondisi tersebut menyebabkan proses analisis secara manual menjadi tidak efisien, memerlukan waktu yang lama, serta berpotensi menghasilkan interpretasi yang subjektif. Selain itu, penggunaan bahasa tidak baku atau slang yang banyak ditemukan pada komentar media sosial dapat memengaruhi kualitas hasil analisis apabila tidak ditangani secara tepat pada tahap prapemrosesan data [3][4][5]. Salah satu isu yang banyak mendapatkan perhatian masyarakat dan menjadi bahan diskusi di media sosial adalah dugaan kasus korupsi di sektor minyak dan gas bumi (migas). Sektor migas merupakan sektor strategis yang memiliki kontribusi besar terhadap perekonomian nasional karena berkaitan langsung dengan penyediaan energi dan penerimaan negara. Pengelolaan sektor ini melibatkan berbagai pihak, termasuk Badan Usaha Milik Negara (BUMN) yang memiliki peran penting dalam kegiatan eksplorasi, produksi, hingga distribusi energi. Salah satu BUMN yang memiliki peran strategis dalam sektor tersebut adalah PT Pertamina (Persero). Sebagai perusahaan energi milik negara, Pertamina memiliki tanggung jawab besar dalam mendukung ketahanan energi nasional sekaligus menjaga tata kelola perusahaan yang transparan, akuntabel, dan berintegritas [6][7].

Namun demikian, sektor migas tidak terlepas dari berbagai permasalahan, salah satunya adalah dugaan tindak pidana korupsi yang melibatkan pihak-pihak tertentu dalam pengelolaannya. Isu korupsi pada sektor strategis seperti migas menjadi perhatian publik karena berpotensi menimbulkan kerugian negara yang besar, mengganggu stabilitas ekonomi, serta menurunkan tingkat kepercayaan masyarakat terhadap institusi pemerintah maupun BUMN. Tingginya perhatian masyarakat terhadap isu tersebut mendorong penyebaran informasi yang masif di berbagai platform digital, termasuk YouTube. Berbagai konten video yang membahas kasus tersebut memunculkan beragam tanggapan dari masyarakat yang tercermin dalam kolom komentar dengan kecenderungan sentimen positif, negatif, maupun netral. Untuk memahami kecenderungan opini publik terhadap suatu isu secara lebih objektif dan terukur, diperlukan suatu metode analisis yang mampu mengolah data komentar dalam jumlah besar. Salah satu pendekatan yang dapat digunakan adalah analisis sentimen. Analisis sentimen merupakan cabang dari text mining yang bertujuan untuk mengidentifikasi, mengekstraksi, dan mengklasifikasikan opini atau emosi yang terkandung dalam data teks. Melalui analisis sentimen, persepsi masyarakat terhadap suatu isu dapat dipetakan secara otomatis sehingga menghasilkan informasi yang lebih sistematis dan representatif. Berbagai penelitian terdahulu menunjukkan bahwa analisis sentimen pada media sosial efektif digunakan untuk mengidentifikasi opini publik terhadap berbagai isu sosial, politik, maupun kebijakan pemerintah [8][9][10].

Meskipun analisis sentimen dapat dilakukan secara manual, pendekatan tersebut memiliki berbagai keterbatasan, terutama dalam aspek efisiensi, konsistensi, dan kemampuan menangani data dalam jumlah besar. Oleh karena itu, diperlukan pendekatan berbasis machine learning yang mampu mengotomatisasi proses klasifikasi sentimen secara lebih cepat dan akurat. Proses tersebut melibatkan beberapa tahapan, mulai dari pengumpulan data, prapemrosesan teks, ekstraksi fitur, hingga pembangunan model klasifikasi yang mampu mengenali pola sentimen dalam data komentar.

Dalam penelitian ini, metode Support Vector Machine (SVM) dipilih sebagai algoritma klasifikasi karena memiliki kemampuan yang baik dalam menangani data berdimensi tinggi seperti data teks. SVM bekerja dengan membentuk hyperplane optimal yang mampu memisahkan data ke dalam kelas-kelas tertentu secara efektif. Karakteristik komentar YouTube yang tidak terstruktur, mengandung bahasa informal, serta memiliki variasi kosakata yang tinggi menjadikan SVM sebagai metode yang relevan untuk digunakan. Penelitian sebelumnya menunjukkan bahwa SVM mampu menghasilkan performa klasifikasi yang lebih baik dibandingkan beberapa metode lain, seperti Naïve Bayes, pada kasus klasifikasi teks [11][12][13][20].

Selain pemilihan algoritma klasifikasi yang tepat, kualitas data juga menjadi faktor penting yang memengaruhi hasil analisis sentimen. Oleh karena itu, penelitian ini menerapkan proses normalisasi kata tidak baku dan pembobotan fitur menggunakan Term Frequency–Inverse Document Frequency (TF-IDF). Teknik normalisasi bertujuan untuk mengubah kata-kata tidak baku menjadi bentuk baku sehingga mengurangi variasi kosakata yang tidak diperlukan, sedangkan TF-IDF digunakan untuk merepresentasikan data teks ke dalam bentuk numerik berdasarkan tingkat kepentingan setiap kata

dalam dokumen. Penelitian terdahulu menunjukkan bahwa kombinasi normalisasi kata dan pembobotan TF-IDF mampu meningkatkan performa klasifikasi SVM pada analisis sentimen media sosial [3][14][15].

Berdasarkan uraian tersebut, penelitian ini dilakukan untuk menganalisis sentimen masyarakat terhadap dugaan kasus korupsi di sektor migas yang melibatkan BUMN melalui komentar pada platform YouTube menggunakan metode Support Vector Machine (SVM). Hasil penelitian diharapkan dapat memberikan gambaran mengenai persepsi publik terhadap isu tersebut sekaligus menjadi referensi dalam pemanfaatan teknik machine learning untuk analisis sentimen pada data media sosial.

TINJAUAN PUSTAKA

Penelitian terdahulu merupakan salah satu tahapan penting dalam penyusunan penelitian yang bertujuan untuk memperoleh referensi yang relevan serta memperkuat landasan teori yang digunakan. Melalui kajian terhadap penelitian sebelumnya, peneliti dapat memahami perkembangan penelitian yang telah dilakukan, metode yang digunakan, serta hasil yang diperoleh sehingga dapat dijadikan sebagai dasar dalam mengembangkan penelitian yang sedang dilakukan. Selain itu, penelitian terdahulu juga berfungsi untuk mengidentifikasi kelebihan dan kekurangan dari penelitian sebelumnya sehingga dapat ditemukan celah penelitian (research gap) yang menjadi dasar dilakukannya penelitian ini. Dalam penelitian ini, peneliti mengkaji beberapa penelitian yang berkaitan dengan analisis sentimen pada media sosial menggunakan pendekatan machine learning dan natural language processing (NLP).

Berdasarkan hasil kajian, penelitian yang dilakukan oleh Nur et al. [3] menunjukkan bahwa algoritma machine learning mampu mengklasifikasikan sentimen komentar YouTube dengan baik. Penelitian oleh Manoppo et al. [16] menggunakan model IndoBERT dan memperoleh akurasi sebesar 84,94% dalam menganalisis sentimen publik terhadap kenaikan PPN 12%. Selanjutnya, Caroline et al. [17] menganalisis sentimen masyarakat terhadap kasus korupsi PT Timah melalui komentar YouTube menggunakan metode Support Vector Machine yang dikombinasikan dengan teknik penyeimbangan data SMOTE dan memperoleh akurasi sebesar 96%. Penelitian Parameswari et al. [18] membuktikan bahwa Support Vector Machine (SVM) memiliki performa yang lebih baik dibandingkan Naïve Bayes dalam analisis sentimen pengguna metaverse dengan akurasi mencapai 78,3%. Sementara itu, Darmayasa et al. [19] menunjukkan bahwa teknik penanganan negasi mampu meningkatkan akurasi klasifikasi sentimen pada algoritma SVM hingga mencapai 89,41%. Meskipun berbagai penelitian tersebut telah berhasil menerapkan analisis sentimen pada berbagai platform media sosial dan topik yang berbeda, penelitian yang secara khusus mengkaji sentimen masyarakat terhadap dugaan kasus korupsi sektor migas melalui komentar YouTube dengan metode Support Vector Machine (SVM) masih terbatas. Oleh karena itu, penelitian ini dilakukan untuk mengisi celah tersebut dengan memanfaatkan komentar YouTube sebagai sumber data dan menerapkan SVM sebagai metode klasifikasi sentimen.

Tabel 1. Ringkasan Penelitian Terdahulu

No	Peneliti & Tahun	Judul Penelitian	Metode	Dataset/Sumber Data	Hasil Penelitian
1	Nur et al. (2025) [3]	<i>Sentiment Analysis of YouTube Comments Using Machine Learning Models</i>	Machine Learning (SVM dan algoritma klasifikasi lainnya)	Komentar YouTube	Model machine learning mampu mengklasifikasikan sentimen positif, negatif, dan netral dengan baik serta menunjukkan efektivitas dalam analisis komentar YouTube.
2	Manoppo et al. (2025) [16]	<i>Analisis Sentimen Publik di Media Sosial terhadap Kenaikan PPN 12% di Indonesia Menggunakan IndoBERT</i>	IndoBERT (Transformer)	X, Instagram, dan TikTok	Menghasilkan akurasi 84,94%, precision 85,6%, recall 84,94%, dan F1-score 84,37%. Sentimen negatif mendominasi opini publik.

3	Caroline et al. (2024) [17]	<i>Analisis Sentimen Masyarakat terhadap Kasus Korupsi PT. Timah Menggunakan Metode Support Vector Machine</i>	Support Vector Machine (SVM) + SMOTE	Komentar YouTube	SVM tanpa SMOTE memperoleh akurasi 89%, sedangkan dengan penerapan SMOTE akurasi meningkat menjadi 96%.
4	Parameswari et al. (2025) [18]	<i>Studi Perbandingan Naïve Bayes dan Support Vector Machine dalam Analisis Sentimen Pengguna Metaverse</i>	Naïve Bayes dan SVM	Data Twitter (X)	SVM memperoleh akurasi 78,3% dan lebih unggul dibandingkan Naïve Bayes yang memperoleh akurasi 72,4%.
5	Darmayasa et al. (2025) [19]	<i>Pengaruh Teknik Penanganan Negasi dalam Analisis Sentimen</i>	Naïve Bayes dan SVM dengan teknik negasi	Ulasan Tokopedia dan Lazada	Teknik penanganan negasi meningkatkan akurasi SVM hingga 89,41% dan terbukti meningkatkan kualitas klasifikasi sentimen.

Tabel 2. Research Gap Penelitian

Aspek	Penelitian Terdahulu	Penelitian Yang Diusulkan
Objek Penelitian	Kebijakan PPN, Korupsi PT Timah, Metaverse, Ulasan Marketplace	Dugaan kasus korupsi sektor migas/BUMN
Platform Data	X, Instagram, YouTube umum, Marketplace	YouTube dengan fokus isu korupsi migas
Metode	IndoBERT, Naïve Bayes, SVM, Machine Learning umum	Support Vector Machine (SVM)
Fokus Analisis	Sentimen terhadap kebijakan, teknologi, dan layanan publik	Sentimen masyarakat terhadap isu korupsi sektor migas
Kontribusi	Analisis sentimen pada berbagai domain	Analisis sentimen publik pada isu korupsi migas berbasis komentar YouTube menggunakan SVM

METODOLOGI

Penelitian ini menggunakan pendekatan kuantitatif dengan metode analisis sentimen berbasis *text mining* dan *machine learning* untuk mengklasifikasikan opini masyarakat pada komentar YouTube terhadap kasus dugaan korupsi migas ke dalam kategori positif, negatif, dan netral. Objek penelitian berupa komentar pengguna YouTube pada video yang membahas isu tersebut, yang merupakan data sekunder hasil *scraping* menggunakan YouTube Data API pada lingkungan Python dan Google Colab. Pengumpulan data dilakukan dengan menentukan kata kunci pencarian, memilih video relevan berdasarkan jumlah penonton dan tingkat interaksi, mengambil komentar dari video terpilih, lalu menyimpannya dalam format CSV untuk dianalisis lebih lanjut.

Penelitian dilaksanakan melalui tujuh tahapan yang saling berurutan, dimulai dari pengumpulan data hingga visualisasi hasil, sebagaimana ditunjukkan pada Gambar 1.



Gambar 1. Diagram Alur Tahapan Penelitian

Tahap *preprocessing* membersihkan dan menyiapkan data melalui *cleaning* (penghapusan URL, emoji, tanda baca, angka, *hashtag*, dan *mention*), *case folding* (penyeragaman ke huruf kecil), *tokenizing* (pemecahan kalimat menjadi token), normalisasi kata tidak baku ke bentuk baku sesuai KBBI (misalnya “gk” → “tidak”, “bgt” → “banget”), *stopword removal* (penghapusan kata tak bermakna seperti “dan”, “yang”, “di”), dan *stemming* ke kata dasar menggunakan Sastrawi (misalnya “mengambil” → “ambil”). Data hasil *preprocessing* kemudian diberi label sentimen ke dalam tiga kategori, yaitu positif (penilaian mendukung), negatif (kritik atau kecaman), dan netral (informatif tanpa emosi tertentu), baik secara manual maupun dengan bantuan kamus sentimen bahasa Indonesia.

Data teks selanjutnya dikonversi menjadi data numerik menggunakan metode *Term Frequency–Inverse Document Frequency* (TF-IDF) yang memberikan bobot pada setiap kata berdasarkan tingkat kemunculannya dalam dokumen, sebagaimana dirumuskan pada Persamaan (1).

$$\text{TF-IDF}(t, d) = \text{TF}(t, d) \times \text{IDF}(t) \quad (1)$$

Dengan TF (*Term Frequency*) menyatakan frekuensi kemunculan kata dalam dokumen dan IDF (*Inverse Document Frequency*) menyatakan tingkat kepentingan kata pada keseluruhan dokumen. Matriks fitur hasil TF-IDF menjadi masukan bagi algoritma *Support Vector Machine* (SVM), yang bekerja dengan mencari *hyperplane* optimal untuk memisahkan kelas sentimen. SVM dipilih karena mampu menangani data berdimensi tinggi, akurat pada data teks, dan tidak mudah mengalami *overfitting*. Data dibagi menggunakan *train-test split* dengan proporsi 80% data latih dan 20% data uji.

Kinerja model dievaluasi menggunakan *confusion matrix* melalui metrik *accuracy*, *precision*, *recall*, dan *F1-score* yang dirumuskan pada Persamaan (2) sampai (5).

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (3)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (4)$$

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (5)$$

Dengan TP, TN, FP, dan FN berturut-turut menyatakan *true positive*, *true negative*, *false positive*, dan *false negative*. Hasil klasifikasi divisualisasikan dalam bentuk diagram batang dan *pie chart* distribusi sentimen, *word cloud* kata yang paling sering muncul, serta *confusion matrix*. Seluruh proses dijalankan menggunakan Python pada Google Colab dengan pustaka Pandas, NumPy, Scikit-Learn, Sastrawi, Matplotlib, Seaborn, dan WordCloud.

HASIL DAN PEMBAHASAN

Pengumpulan dan Praproses Data

Implementasi sistem dilakukan pada Google Colab menggunakan bahasa Python. Opini publik dikumpulkan melalui teknik *crawling* terhadap komentar video YouTube yang membahas kasus dugaan korupsi minyak dan gas bumi (migas) dengan memanfaatkan pustaka *google-api-python-client* dan *YouTube Data API v3*. Data mentah distrukturkan ke dalam *DataFrame* berisi atribut *Comment ID*, *Username*, *Comment*, dan *Published At*, lalu disimpan dalam berkas CSV.

Komentar media sosial yang tidak terstruktur dan penuh derau (*noise*) memerlukan serangkaian praproses, yaitu *cleaning*, *case folding*, *normalization*, *tokenizing*, *stopword removal*, dan *stemming*. Tahapan ini menurunkan dimensi fitur sekaligus mempertajam makna kata. Perbandingan kondisi data sebelum dan sesudah praproses disajikan pada Tabel 3, sedangkan pengaruhnya secara visual ditampilkan pada Gambar 1.

Tabel 3. Contoh Hasil Tahapan Praproses Data Teks

No	Teks Mentah (Raw Comment)	Teks Hasil Preprocessing
1	1 orang mudah sedang paletina bisa merdeka	orang mudah palestina merdeka
2	si nabl anjlng 🤔🤔, kadrn suoe ngehe Ajaran bejad	nab anjing kadrn ajar bejat
3	Tidak mungkin bisa. Ketangkap. Paling ketemu di...	tidak mungkin tangkap ketemu
4	kenapa takut sama riza halid dulu klu ada kasus	takut riza chalid kasus

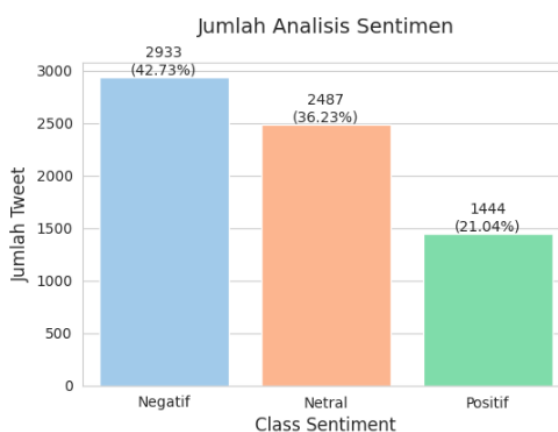
Sebelum pembersihan, *word cloud* didominasi kata fungsional seperti “yg”, “dan”, dan “nya”. Setelah praproses, kata substantif seperti “negara”, “rakyat”, “koruptor”, dan “hukum” muncul dominan, membuktikan derau bahasa berhasil dihilangkan dan esensi opini publik terlihat jelas.



Gambar 2. WordCloud Sebelum dan Sesudah Praproses

Pelabelan dan Karakteristik Sentimen

Pelabelan dilakukan otomatis dengan pendekatan *lexicon-based*, yaitu mencocokkan setiap kata dengan kamus sentimen bahasa Indonesia bermuatan polaritas positif dan negatif untuk menetapkan tiga kelas: negatif, netral, dan positif. Dari 6.864 dokumen, hasil pelabelan menunjukkan sentimen negatif 2.933 dokumen (42,73%), netral 2.487 dokumen (36,23%), dan positif 1.444 dokumen (21,04%) seperti pada Gambar 2, yang mencerminkan tingginya kecenderungan kritis publik.

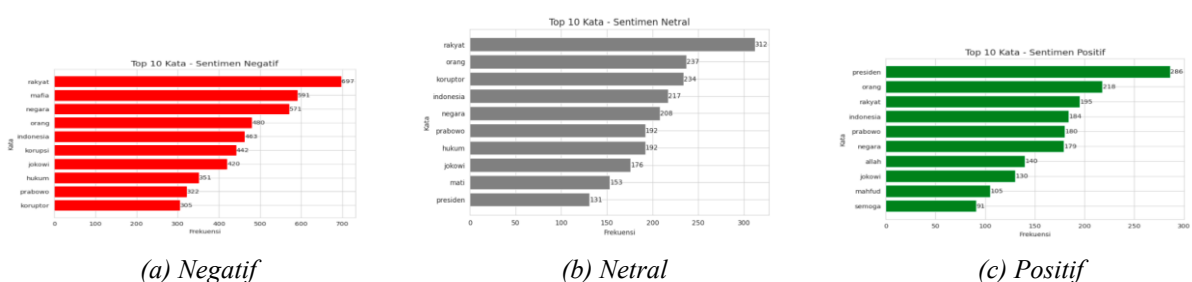


Gambar 3. Distribusi Kelas Sentimen Hasil Pelabelan Lexicon-Based

Pemetaan *word cloud* per kelas (Gambar 4) memperlihatkan perbedaan diksi: kelas negatif didominasi kosakata kritik seperti “mafia” dan “korupsi”; kelas netral cenderung objektif dengan kata benda umum seperti “rakyat” dan “negara”; sedangkan kelas positif menonjolkan ungkapan spiritual seperti “allah”, “semoga”, dan “dukung”.

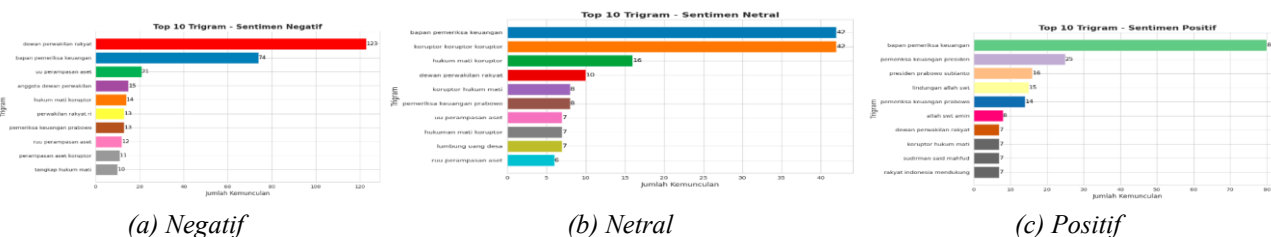


Analisis frekuensi kata tunggal (*unigram*) per kelas (Gambar 4) memberi bukti kuantitatif. Kelas negatif dipuncaki “rakyat” (697), “mafia” (591), dan “negara” (572), menandakan persepsi benturan kepentingan publik dengan kartel energi dan desakan hukum (“hukum” 351). Kelas netral didominasi “rakyat” (313), “orang” (237), dan “koruptor” (234) tanpa polaritas. Kelas positif dipuncaki “presiden” (399), “rakyat” (278), dan “prabowo” (273), dengan muatan spiritual “allah” (209) dan “semoga” (125).

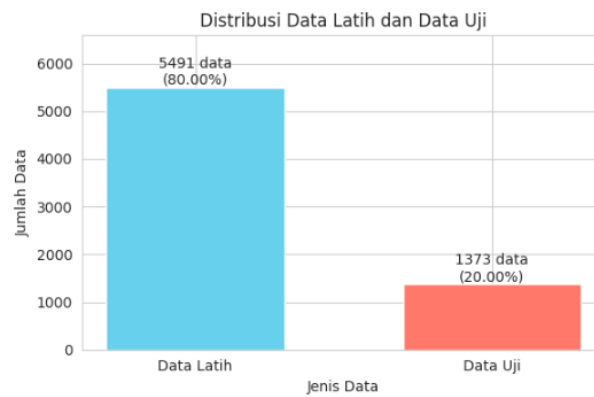


Analisis *trigram* (Gambar 5) memperkaya konteks. Kelas negatif didominasi “dewan perwakilan rakyat” (123) dan “badan pemeriksa keuangan” (74), dipertegas “uu perampasan aset” (21) dan “hukum mati koruptor” (14). Kelas netral memuat pengulangan “koruptor koruptor koruptor” (42) khas salin-tempel berita. Kelas positif menampilkan “badan pemeriksa keuangan” (80), “presiden prabowo subianto” (16), dan frasa spiritual “lindungan allah swt” (15).

Pembagian Data dan Perbandingan Kernel SVM



Korpus 6.864 dokumen dibagi dengan *train_test_split* (*scikit-learn*) menjadi 80% data latih (5.491 dokumen) dan 20% data uji (1.373 dokumen) seperti pada Gambar 7. Parameter *stratify* diaktifkan agar distribusi tiap kelas seimbang pada kedua subset, mencegah bias kelas mayoritas dan kebocoran data (*data leakage*).

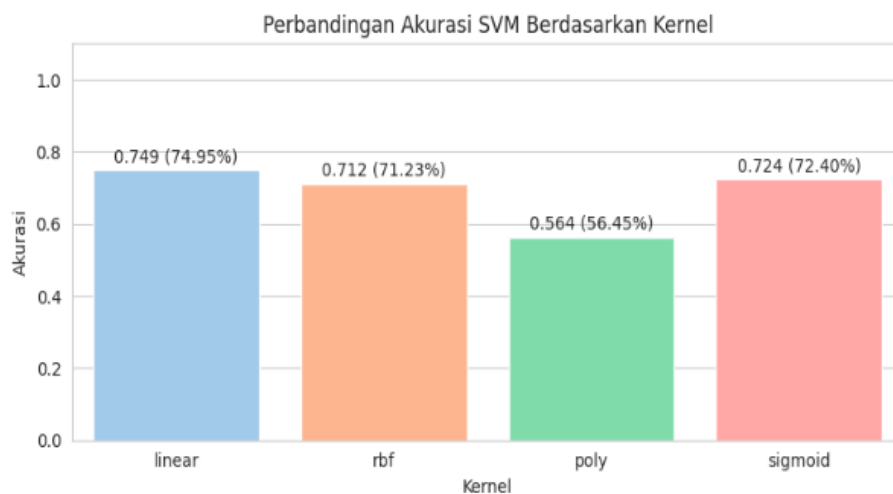


Gambar 7. Pembagian Data Latih dan Data Uji

Fitur tekstual diekstraksi dengan pembobotan *Term Frequency-Inverse Document Frequency* (TF-IDF), lalu diklasifikasikan menggunakan *Support Vector Machine* (SVM) dengan empat kernel: *Linear*, *Sigmoid*, *RBF*, dan *Polynomial*. Ringkasan performa disajikan pada Tabel 4 dan Gambar 8.

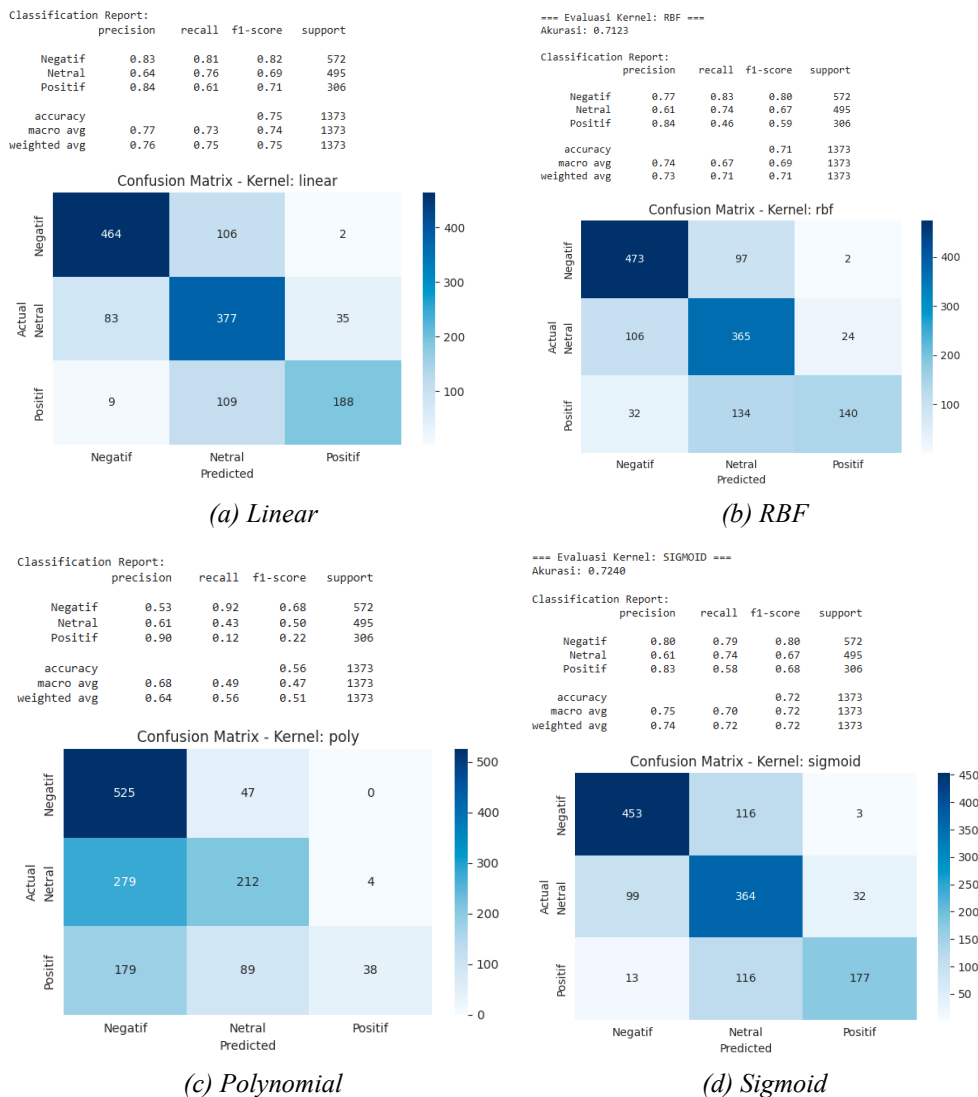
Tabel 4. Perbandingan Performa Kernel SVM pada Data Uji

Kernel SVM	Accuracy	Precision	Recall	F1-Score
Linear	74,95%	0,75	0,75	0,75
Sigmoid	72,40%	0,72	0,72	0,72
RBF	71,23%	0,72	0,71	0,71
Polynomial	56,45%	0,58	0,56	0,51



Gambar 8. Perbandingan Akurasi Kernel SVM

Kernel *Linear* mencatat performa tertinggi (akurasi 74,95%, *weighted F1-score* 0,75), diikuti *Sigmoid* (72,40%) dan *RBF* (71,23%), sedangkan *Polynomial* terendah (56,45%). Keunggulan ini mengonfirmasi matriks TF-IDF bersifat *linearly separable*, sehingga garis pemisah linear sudah cukup tanpa pemetaan non-linear yang mahal secara komputasi. *Confusion matrix* keempat kernel ditampilkan pada Gambar 9.



Gambar 9. Confusion Matrix Empat Kernel SVM

Pada kernel *Linear*, kelas negatif terklasifikasi paling baik (*precision* 0,83, *recall* 0,81; 464 dari 572 dokumen tepat), namun 109 dokumen positif keliru menjadi netral akibat kosakata subjek bersama seperti “presiden”, “prabowo”, dan “jokowi”. Kernel non-linear kurang sesuai untuk fitur TF-IDF yang renggang: *RBF* menurun pada *recall* positif (0,46) dan *Polynomial* mengalami *underfitting* parah (*recall* positif 0,12), sedangkan *Sigmoid* stabil di peringkat kedua. Secara keseluruhan, kombinasi TF-IDF dengan SVM kernel *Linear* merupakan konfigurasi paling andal; pola salah klasifikasi positif–netral menyiratkan perlunya integrasi fitur *n-gram* yang lebih kaya pada penelitian selanjutnya.

KESIMPULAN DAN SARAN

Berdasarkan hasil penelitian yang telah dilakukan mengenai analisis sentimen komentar YouTube terhadap kasus dugaan korupsi migas menggunakan algoritma Support Vector Machine (SVM), dapat disimpulkan bahwa metode SVM yang dikombinasikan dengan pembobotan fitur Term Frequency-Inverse Document Frequency (TF-IDF) mampu mengklasifikasikan opini masyarakat ke dalam kategori sentimen positif, negatif, dan netral secara efektif. Penggunaan TF-IDF berhasil mengubah data teks yang tidak terstruktur menjadi representasi numerik sehingga dapat diproses oleh model klasifikasi. Karakteristik data komentar YouTube yang berdimensi tinggi dan memiliki tingkat keragaman kosakata yang besar dapat ditangani dengan baik oleh algoritma SVM. Hasil pengujian menunjukkan bahwa pemilihan fungsi kernel berpengaruh terhadap performa klasifikasi yang dihasilkan. Dari empat kernel yang diuji, yaitu linear, sigmoid, Radial Basis Function (RBF), dan polynomial, kernel linear menghasilkan performa terbaik dengan tingkat akurasi sebesar 74,95% dan weighted F1-score sebesar 0,75. Hasil tersebut menunjukkan bahwa pola sebaran data sentimen pada komentar YouTube terkait kasus dugaan korupsi migas cenderung dapat dipisahkan secara linear sehingga tidak memerlukan transformasi fitur yang kompleks untuk memperoleh hasil klasifikasi yang optimal. Sementara itu, kernel

non-linear menunjukkan performa yang lebih rendah dibandingkan kernel linear. Kernel sigmoid memperoleh akurasi sebesar 72,40% dengan F1-score 0,72, sedangkan kernel RBF menghasilkan akurasi sebesar 71,23% dengan nilai recall yang masih rendah pada kelas sentimen positif. Adapun kernel polynomial memberikan performa terendah dengan akurasi sebesar 56,45% dan F1-score sebesar 0,51. Temuan ini menunjukkan bahwa kompleksitas model yang terlalu tinggi tidak selalu meningkatkan performa klasifikasi, terutama pada data teks media sosial yang memiliki karakteristik bahasa informal dan distribusi data yang tidak merata.

Berdasarkan hasil penelitian dan keterbatasan yang ditemukan selama proses analisis, penelitian selanjutnya disarankan untuk mengembangkan proses ekstraksi fitur dengan menerapkan kombinasi n-gram, seperti unigram, bigram, dan trigram. Penggunaan n-gram dapat membantu model memahami konteks antar kata secara lebih baik sehingga makna kalimat dapat direpresentasikan secara lebih akurat dibandingkan hanya menggunakan unigram. Dengan demikian, potensi kesalahan klasifikasi akibat hilangnya konteks kalimat dapat diminimalkan. Selain itu, penelitian berikutnya dapat menerapkan teknik penanganan ketidakseimbangan data (imbalanced data) seperti Synthetic Minority Over-sampling Technique (SMOTE) atau metode serupa. Teknik tersebut diharapkan mampu meningkatkan kemampuan model dalam mengenali kelas sentimen minoritas, khususnya sentimen positif, sehingga distribusi prediksi menjadi lebih seimbang dan nilai recall pada setiap kelas dapat meningkat. Penelitian selanjutnya juga dapat melakukan perbandingan antara algoritma Support Vector Machine dengan metode machine learning maupun deep learning lainnya, seperti Random Forest, Long Short-Term Memory (LSTM), Bidirectional Encoder Representations from Transformers (BERT), atau IndoBERT. Perbandingan tersebut diharapkan dapat memberikan gambaran yang lebih komprehensif mengenai metode yang paling efektif dalam menganalisis sentimen masyarakat terhadap isu-isu publik yang berkembang di media sosial, khususnya pada platform YouTube.

DAFTAR PUSTAKA

- [1] R. Singha, S. Das, K. Das, and M. Hasan, "Sentiment analysis of YouTube comments using machine learning algorithm," *ieeexplore.ieee.org* R Singha, S Das, K Das, M Hasan 2024 IEEE Int. Conf. Comput. Appl. and, 2024, *ieeexplore.ieee.org*, 2024, doi: 10.1109/COMPAS60761.2024.10797211.
- [2] F. Syofiani, S. Alam, and M. I. Sulisty, "Analisis Sentimen Penilaian Masyarakat terhadap Childfree Berdasarkan Komentar di YouTube Menggunakan Algoritma Naïve Bayes," *J. Teknol. Inform. dan Komput.*, vol. 9, no. 2, pp. 688–703, 2023, doi: 10.37012/jtik.v9i2.1661.
- [3] A. Nur, A. Saputra, R. E. Saputro, and S. Saputra, "Enhancing sentiment analysis accuracy using SVM and slang word normalization on YouTube comments," *jurnal.polgan.ac.id* ANA Saputra, RE Saputro, DIS Saputra Sinkron J. dan Penelit. Tek. Inform. 2025 • jurnal.polgan.ac.id, vol. 9, no. 2, pp. 687–699, Apr. 2025, doi: 10.33395/SINKRON.V9I2.14613.
- [4] I. Putra, S. Salmon, K. K.-B. of C. Science, and undefined 2025, "Sentimen Masyarakat Terhadap Kenaikan Bahan Bakar Minyak pada Media Sosial Youtube dengan Metode K-Nearest Neighbor dan Support Vector Machine," *ejournal.itn.ac.id*, vol. 6, no. 1, pp. 445–457, 2025, doi: 10.47065/bulletincsr.v6i1.883.
- [5] R. I. Tarecha, F. Wahyudi, and U. M. Jannah, "Penanganan Negasi dalam Analisa Sentimen Bahasa Indonesia," *J. Sist. Inf. dan Inform. (JUSIFOR)*, vol. 1, no. 1, pp. 51–58, 2022, doi: 10.33379/jusifor.v1i1.1276.
- [6] A. Saputra, I. Nurdiani, ... U. N.-J. digital: J., and undefined 2026, "Sentimen Masyarakat Terhadap Penanganan Banjir Bandang di Pulau Sumatra Berdasarkan Komentar Youtube Menggunakan Metode Support Vector Machine ...," *indojurnal.com*, doi: 10.63822/8d1vka43.
- [7] D. Darwis, E. S. Pratiwi, and A. F. O. Pasaribu, "Penerapan Algoritma SVM untuk Analisis Sentimen pada Data Twitter Komisi Pemberantasan Korupsi Republik Indonesia," *Edutic - Sci. J. Informatics Educ.*, vol. 7, no. 1, pp. 1–11, 2020.
- [8] U. Jaya, "KRITIK TERHADAP PEMERINTAHAN JOKO WIDODO PADA PENGUJUNG MASA JABATAN DALAM MAJALAH TEMPO (Analisis Semiotika Roland Barthes)," 2025, Accessed: Jun. 13, 2026. [Online]. Available: <https://digilib.unila.ac.id/90592/>
- [9] A. Setiawan, R. S.-E. Jurnal, and undefined 2024, "Analisis Sentimen Ibu Kota Nusantara menggunakan Algoritma Support Vector Machine dan Naïve Bayes," *e-journal.hamzanwadi.ac.id*, vol. 8, no. 1, pp. 183–192, 2024, doi: 10.29408/edumatic.v8i1.25667.
- [10] S. Styawati, N. Hendrastuty, and A. R. Isnain, "Analisis Sentimen Masyarakat terhadap Program Kartu Prakerja pada Twitter dengan Metode Support Vector Machine," *J. Inform. J. Pengemb. IT*, vol. 6, no. 3, pp. 150–155, 2021, doi: 10.30591/jpit.v6i3.2870.

- [11] J. Saputra, L. Maryani, Rahmadden, D. Wulandari, and W. Eka, "Analisis Performa Naive Bayes dan SVM terhadap Sentimen Teks Media Sosial dengan Word2Vec dan SMOTE," *J. INSTEK (Inform. Sains dan Teknol.)*, vol. 10, no. 1, pp. 143–155, 2025.
- [12] Normah, B. Rifai, S. Vambudi, and R. Maulana, "Analisa Sentimen Perkembangan Vtuber Dengan Metode Support Vector Machine Berbasis SMOTE," *J. Tek. Komput. AMIK BSI*, vol. 8, no. 2, pp. 174–180, 2022, doi: 10.31294/jtk.v4i2.
- [13] T. Maulana, B. Adhiyaksa, R. A. Fauzi, R. I. Agustin, S. A. Azhaar, and Rohana, "Komparasi Algoritma Naive Bayes dan SVM untuk Analisis Sentimen Twitter Korupsi Bansos Beras Masa Pandemi," *J. Inform. dan Tek. Elektro Terap. (JITET)*, vol. 12, no. 2, pp. 912–918, 2024, doi: 10.23960/jitet.v12i2.4020.
- [14] F. Rahmadayana and Y. Sibaroni, "Sentiment Analysis of Work from Home Activity using SVM with Randomized Search Optimization," *J. RESTI (Rekayasa Sist. dan Teknol. Inf.)*, vol. 5, no. 5, pp. 936–942, 2021, doi: 10.29207/resti.v5i5.3457.
- [15] M. Goswami, ... P. S. I. S. P. of E., and undefined 2021, "A comparative analysis of sentiment analysis using RNN-LSTM and logistic regression," *Springer*, vol. 740 LNEE, pp. 165–174, 2021, doi: 10.1007/978-981-33-6393-9_18.
- [16] M. R. Manoppo *et al.*, "Analisis Sentimen Publik di Media Sosial terhadap Kenaikan PPN 12% di Indonesia Menggunakan IndoBERT," *J. Kecerdasan Buatan dan Teknol. Inf.*, vol. 4, no. 2, pp. 152–163, 2025, doi: 10.69916/jkbt.v4i2.322.
- [17] F. Caroline, R. G. S. Budi, and M. E. A. Rivan, "Analisis Sentimen Masyarakat terhadap Kasus Korupsi PT. Timah Menggunakan Metode Support Vector Machine," *J. Ilmu Komput. dan Inform. (JIKI)*, vol. 4, no. 1, pp. 43–50, 2024, doi: 10.54082/jiki.141.
- [18] S. D. Parameswari, M. Lubis, S. Suakanto, Y. Z. Ramadhan, R. N. Amanah, and R. A. Dila, "Studi Perbandingan Naive Bayes dan Support Vector Machine (SVM) dalam Analisis Sentimen Pengguna Metaverse," *J. Teknol. dan Manaj. Ind. Terap.*, vol. 4, no. 3, pp. 1059–1065, 2025, doi: 10.55826/jtmit.v4i3.1122.
- [19] I. N. O. Darmayasa, N. A. Sanjaya ER, I. G. A. G. A. Kadyanan, and A. A. I. N. E. Karyawati, "Pengaruh Teknik Penanganan Negasi dalam Analisis Sentimen," *J. Teknol. Inf. dan Ilmu Komput. (JTIK)*, vol. 12, no. 2, pp. 275–282, 2025, doi: 10.25126/jtiik.2025129079.
- [20] H. Setiawan, E. Utami, and S. Sudarmawan, "Analisis Sentimen Twitter Kuliah Online Pasca Covid-19 Menggunakan Algoritma Support Vector Machine dan Naive Bayes," *J. Komtika (Komputasi dan Informatika)*, vol. 5, no. 1, pp. 43–51, 2021, doi: 10.31603/komtika.v5i1.5189.