Jurnal Dinamika Informatika
Volume 15, No 1, Maret 2026
ISSN 1978-1660 : 29-43
ISSN *online* 2549-8517

Analisis Sentimen Video YouTube KOMPASTV "Pajak Cekik Rakyat, Tunjangan DPR Naik" Menggunakan Naive Bayes & SVM

Novi Eka Rahmawati[1], Rahmatul Ummah[2], Harun Al Rosyid[3]

[1,2,3]Program Studi Pendidikan Teknologi Informasi, Fakultas Teknik, Universitas Negeri Surabaya

e-mail: **[1]novi.22044@mhs.unesa.ac.id, [2]rahmatul.22059@mhs.unesa.ac.id, [3]harunrosyid@unesa.ac.id**

Intisari

Penelitian ini bertujuan untuk menganalisis sentimen publik terhadap video YouTube KOMPASTV berjudul "Pajak Cekik Rakyat, Tunjangan DPR Naik" dengan menggunakan metode Naive Bayes dan Support Vector Machine (SVM). Topik ini dipilih karena video tersebut membahas isu-isu sensitif terkait kebijakan pajak dan tunjangan DPR, yang dapat mempengaruhi persepsi masyarakat. Metode penelitian yang digunakan adalah analisis sentimen berbasis machine learning, dengan pengumpulan data melalui YouTube Data API dan web scraping. Komentar yang terkumpul kemudian dianalisis menggunakan teknik preprocessing seperti stopword removal, tokenisasi, dan stemming. Model klasifikasi yang diterapkan pada data adalah Naive Bayes dan SVM, dengan evaluasi menggunakan metrik akurasi, presisi, recall, dan F1-score. Hasil penelitian menunjukkan bahwa model SVM memiliki akurasi lebih tinggi dibandingkan Naive Bayes dalam mengklasifikasikan sentimen positif, negatif, dan netral. Penelitian ini memberikan wawasan mengenai bagaimana algoritma machine learning dapat digunakan untuk memahami dinamika opini publik melalui komentar-komentar di media sosial.

Kata kunci—analisis sentimen, Naive Bayes, SVM, YouTube, pajak

Abstract

This study aims to analyze public sentiment toward the KOMPASTV YouTube video titled "Pajak Cekik Rakyat, Tunjangan DPR Naik" using Naive Bayes and Support Vector Machine (SVM) methods. This topic was chosen because the video discusses sensitive issues related to tax policies and DPR allowances, which can influence public perceptions. The research method used is sentiment analysis based on machine learning, with data collected through the YouTube Data API and web scraping. The collected comments were then analyzed using preprocessing techniques such as stopword removal, tokenization, and stemming. The classification models applied to the data are Naive Bayes and SVM, with evaluation using accuracy, precision, recall, and F1-score metrics. The results show that the SVM model outperforms Naive Bayes in classifying positive, negative, and neutral sentiments. This research provides insights into how machine learning algorithms can be used to understand the dynamics of public opinion through social media comments.

Keywords*—sentiment analysis, Naive Bayes, SVM, YouTube, tax*

PENDAHULUAN

Platform media sosial, khususnya YouTube, telah berkembang menjadi alat penting dalam membentuk opini publik. YouTube, meskipun pada dasarnya adalah platform berbagi video, juga memiliki elemen sosial yang memungkinkan interaksi antar pengguna melalui komentar, langganan saluran, dan berbagi video [1]. Sebagai salah satu platform dengan jumlah pengguna terbesar, YouTube berperan signifikan dalam menyebarkan informasi yang dapat mempengaruhi pandangan masyarakat terhadap isu-isu tertentu. Dalam konteks ini, analisis sentimen terhadap video yang membahas topik-topik sensitif seperti kebijakan pajak dan tunjangan DPR menjadi sangat relevan untuk memahami dinamika opini publik [2].

Topik-topik kontroversial yang sering dibahas dalam media, termasuk kebijakan pemerintah, mendapatkan perhatian besar di YouTube karena audiens yang luas dan keberagaman opini yang muncul di kolom komentar. Video seperti "Pajak Cekik Rakyat, Tunjangan DPR Naik" di kanal KOMPASTV, misalnya, dapat memicu beragam respons dari masyarakat yang menunjukkan persepsi mereka terhadap kebijakan publik yang sedang dibahas. Dengan jumlah penayangan yang signifikan, video seperti ini dapat memengaruhi persepsi sosial, dan oleh karena itu, penting untuk menganalisis sentimen yang berkembang di sekitar video tersebut untuk mendapatkan gambaran yang lebih jelas mengenai opini publik terhadap isu tersebut [3].

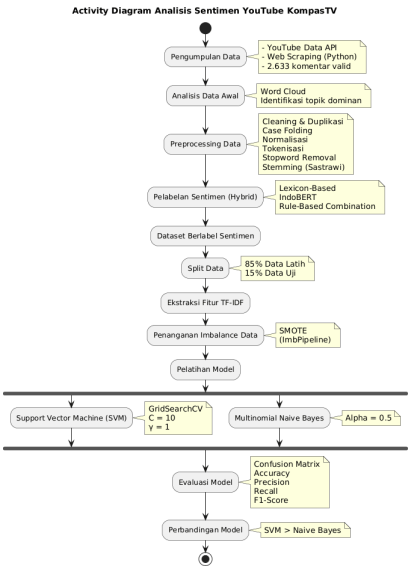
Penelitian ini bertujuan untuk menganalisis sentimen publik terhadap video yang membahas isu-isu kontroversial dengan menggunakan metode Naive Bayes dan SVM. Tujuan utamanya adalah untuk memberikan wawasan mengenai bagaimana publik merespons topik sensitif seperti pajak dan tunjangan DPR, serta bagaimana hal ini bisa digunakan untuk memahami dampak sosial dari video tersebut. Penelitian ini terbatas pada komentar-komentar di video YouTube yang dimuat di kanal KOMPASTV, dengan fokus pada data yang tersedia seperti jumlah penayangan dan komentar yang diterima. Alat analisis yang digunakan akan membantu dalam menggali pandangan masyarakat terkait video tersebut dan memberikan gambaran yang lebih objektif mengenai sentimen yang ada [4].

METODE PENELITIAN

Penelitian ini mengaplikasikan pendekatan kuantitatif komputasional dengan menggunakan metode analisis sentimen berbasis pembelajaran mesin. Algoritma yang digunakan adalah *Naive Bayes* dan *Support Vector Machine* (SVM) yang digunakan untuk menentukan kategori sentimen komentar pengguna terhadap video YouTube KOMPASTV. Tujuan utama metode ini yaitu mengukur performa kedua algoritma pada dataset komentar yang sama dan menentukan algoritma mana yang lebih akurat dalam mengklasifikasi sentimen positif, netral, atau negatif.

1. Alur Kerja Penelitian

Jurnal Dinamika Informatika
Volume 15, No 1, Maret 2026
ISSN 1978-1660 : 29-43
ISSN *online* 2549-8517



Gambar 1. Alur Kerja Penelitian

Alur kerja analisis sentimen pada penelitian ini digambarkan dalam bentuk *Activity Diagram* yang menunjukkan tahapan proses secara sistematis, meliputi: (1) Pengumpulan data menggunakan YouTube Data API dan web scraping dengan Python untuk memperoleh komentar-komentar yang relevan. (2) Analisis data, mencakup pembuatan *word cloud*, cleaning text, dan *tokenization*. (3) Preprocessing data, meliputi Stopword Removal dan Stemming. (4) Ekstraksi fitur menggunakan teknik TF-IDF. (5) Pembagian dataset menjadi data latih dan data uji. (6) Imbalance data menggunakan metode SMOTE (*Synthetic Minority Over-sampling Technique*). (7) Pelatihan model menggunakan model *hybrid*, yaitu mengombinasikan metode lexicon-based dan model IndoBERT. (6) Model yang dibangun diuji menggunakan dua metode, yaitu Support Vector Machine (SVM) dan Multinomial Naive Bayes. (8) Evaluasi model menggunakan Confusion Matrix untuk mengukur *accuracy*, *precision*, *recall*, dan *F1-score*.

2. Pengumpulan Data

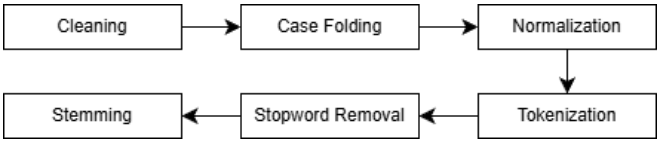
Penelitian dimulai dengan tahap pengumpulan data komentar dari kanal YouTube KompasTV. Pengambilan data dilakukan menggunakan YouTube Data API dan bahasa pemrograman Python, sehingga diperoleh sejumlah 2.633 komentar valid untuk dianalisis. Data API memungkinkan peneliti secara terstruktur mendapatkan komentar yang relevan berdasarkan video target penelitian, tanpa harus melakukan scraping manual [5].

3. Analisis Data Awal

Analisis awal dilakukan untuk melihat pola dominan dari kata-kata yang sering muncul dalam komentar. Teknik *Word Cloud* digunakan untuk mendapatkan gambaran umum distribusi kata dalam korpus. Hal ini berguna untuk mengidentifikasi topik umum sebelum dilakukan preprocessing lebih lanjut.

4. Preprocessing Data

Sebelum teks dimodelkan, dilakukan beberapa tahap preprocessing, meliputi:



Gambar 2. Tahap Preprocessing Data

Tahapan ini dilakukan untuk mengurangi *noise*, menyamarkan format teks, serta memperkecil variasi kata. Preprocessing ini serupa dengan prosedur tekstual di berbagai studi klasifikasi teks menggunakan TF-IDF dan SVM/Naive Bayes [6].

5. Pelabelan Sentimen

   Pelabelan sentimen dilakukan menggunakan pendekatan hibrida, yaitu mengombinasikan metode lexicon-based dan model IndoBERT. Metode lexicon digunakan untuk menghitung kecenderungan sentimen berdasarkan kamus kata positif dan negatif, sedangkan IndoBERT digunakan untuk memahami konteks kalimat secara lebih mendalam [7]. Hasil dari kedua metode tersebut kemudian digabungkan menggunakan aturan berbasis rule-based combination untuk menghasilkan label sentimen akhir berupa positif, negatif, atau netral.

6. Pembagian Dataset

   Dataset yang telah memiliki label sentimen selanjutnya dibagi menjadi dua bagian, yaitu *train data* sebesar 85% dan *test data* sebesar 15%. Split data ini bertujuan untuk melatih model klasifikasi serta menguji performa model pada data yang belum pernah dilihat sebelumnya.

7. Ekstraksi Fitur

   Pada tahap ini, data teks diubah ke dalam bentuk numerik agar dapat diproses oleh algoritma machine *learning*. Metode yang digunakan adalah *Term Frequency–Inverse Document Frequency* (TF-IDF), yang merepresentasikan tingkat pentingnya suatu kata dalam sebuah dokumen dibandingkan dengan keseluruhan dataset.

   Pembobotan TF IDF menggunakan perhitungan,

$$W_{t,d} = TF_{t,d} \times IDF_t \quad [6]$$

   Dimana IDF_t dihitung dengan

$$IDF_t = \log\left(\frac{N}{DF_t}\right) \quad [6]$$

Keterangan:

$W_{t,d}$: Bobot kata t dalam dokumen d

$TF_{t,d}$: Frekuensi kemunculan kata t dalam dokumen d

N : Total jumlah dokumen.

IDF_t : Jumlah dokumen yang mengandung kata t.

8. Imbalance Data

   Karena jumlah data pada setiap kelas sentimen tidak seimbang, dilakukan penanganan imbalance data menggunakan metode *Synthetic*

Jurnal Dinamika Informatika
Volume 15, No 1, Maret 2026
ISSN 1978-1660 : 29-43
ISSN *online* 2549-8517

Minority Over-sampling Technique (SMOTE)[8]. Metode ini menghasilkan data sintesis pada kelas minoritas agar distribusi data menjadi lebih seimbang. Proses SMOTE diterapkan menggunakan ImbPipeline untuk memastikan bahwa oversampling hanya dilakukan pada data latih.

$$x_{new} = x + \lambda \times (x_{nn} - x) \quad [9]$$

Keterangan:
x_{new} : Data sintetis baru.
x : Vektor fitur data minoritas asli.
x_{nn} : Salah satu tetangga terdekat (k-nearest neighbor) dari x.
λ : Angka acak antara 0 dan 1.

9. Pelatihan Model Klasifikasi

Tahap pelatihan model dilakukan menggunakan dua algoritma klasifikasi, yaitu *Support Vector Machine* (SVM) dan *Multinomial Naive Bayes*.

a. Metode Naive Bayes

$$P(c|d) \propto P(c) \prod_{k=i}^{n_d} P(t_k|c) \quad [10]$$

Keterangan:
c : kelas
d : dokumen
t_k : token/term ke-k dalam dokumen
n_d : jumlah token dalam dokumen
$P(c)$: prior kelas
$P(t_k|c)$: probabilitas term tkt_ktk muncul pada kelas c

b. Metode Support Vector Machine (SVM)
Rumus Hyperplane terbaik: Garis pemisah (Decision Boundary)

$$f(x) = w^T \phi(x) + b = 0 \quad [10]$$

Rumus keputusan kelas

$$y(x) = sign\left(\sum_{i=1}^{n} \alpha_i y_i K(x_i, x) + b\right) \quad [10]$$

Menggunakan Kernel RBF pada kode optimasi , dengan fungsi kernel

$$K(x_i, x) = \exp\left(-\gamma ||x_i - x||^2\right) \quad [10]$$

Keterangan:
W : Vektor bobot (weight).
b : Bias.
α_i : Lagrange multipliers.
γ : Parameter gamma (dari kode GridSearch).

10. Evaluasi Model

Evaluasi model dilakukan untuk mengukur tingkat keberhasilan model dalam mengklasifikasikan sentimen komentar. Confusion matrix dan metrik evaluasi digunakan sebagai metode evaluasi yang berupa akurasi, *precision*, *recall*, dan F1-score. Hasil evaluasi ini memberikan gambaran kinerja masing-masing model secara kuantitatif.

Akuransi (Accurancy):

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad [8]$$

Presisi (Precision):

$$Precision = \frac{TP}{TP + FP} \quad [8]$$

Recall (Sensitivitas):

$$Recall = \frac{TP}{TP + FN} \quad [8]$$

F1-Score:

$$F1\text{-}Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad [8]$$

Keterangan:

TP (*True Positive*) : Data positif diprediksi positif.
TN (*True Negative*) : Data negatif diprediksi negatif.
FP (*False Positive*) : Data negatif diprediksi positif.
FN (*False Negative*) : Data positif diprediksi negatif.

11. Perbandingan Model

Perbandingan antara model *Support Vector Machine* dan *Multinomial Naive Bayes* dilakukan berdasarkan hasil evaluasi. Hasil perbandingan menunjukkan bahwa model SVM memiliki performa yang lebih baik dibandingkan *Naive Bayes* untuk mengklasifikasikan sentimen komentar YouTube KompasTV.

HASIL DAN PEMBAHASAN

1. Pengumpulan Data

Pengumpulan data menggunakan data sekunder yang bersumber dari respons warganet pada video YouTube KompasTV berjudul "Pajak Cekik Rakyat, Tunjangan DPR Naik". Proses akuisisi data dijalankan menggunakan YouTube Data API serta teknik web scraping berbasis bahasa pemrograman Python dan library terkait. Dari total 2.651 komentar yang terunduh, dilakukan tahap pembersihan (data cleaning) untuk mengeliminasi duplikasi. Hasil akhirnya, tersisa 2.633 data valid yang disimpan dalam format .csv. Sampel hasil pengumpulan data tersebut dapat dilihat pada Tabel 1 berikut.

Jurnal Dinamika Informatika
Volume 15, No 1, Maret 2026
ISSN 1978-1660 : 29-43
ISSN *online* 2549-8517

Table 1. Contoh Data Awal

No	Komentar	Jumlah Like	Waktu Unggah
1	Jgn suka membodohi rakyat pak, gaji nggak naik, tunjangan itu sama juga dengan gaji, apakah tunjangan itu dari daun 😂	0	2025-09-01 12:20:01
2	"Menteri keuangan terbalik, kasihan masa kasihan pada anggota DPR daripada ke rakyat. Salah lihat menteri keuangan."	0	2025-09-01 11:21:20
3	Rakyat menanggung pajak sendiri, sementara DPR dengan gaji besar dan berbagai tunjangan justru ditanggung negara pajaknya. Bagaimana rakyat bisa makmur jika demikian? Padahal mereka bukan siapa-siapa tanpa rakyat.	0	2025-09-01 06:32:49
4	Itu aturannya siapa, kok pintar sekali ya	0	2025-08-308:37:40

2. Analisis Data Awal

Analisis data umumnya digunakan untuk melihat karakteristik dari dataset sebelum memasuki tahapan pemrosesan lebih lanjut. Tahapan ini meliputi pengamatan jumlah data dan kecenderungan topik dari keseluruhan komentar yang sedang dibahas oleh masyarakat. Analisis pada penelitian ini menggunakan word cloud sebagai modelnya yaitu representasi grafis dari data teks yang menonjolkan kata - kata populer berdasarkan frekuensi kemunculan katanya.

Gambar 3. Word Cloud

3. Preprocessing Data

Proses ini memiliki tujuan untuk *cleaning text* agar dapat diproses pada tahap klasifikasi.

a. Cleaning dan Penghapusan Duplikasi

Proses ini melibatkan penghapusan tanda baca dan emoji yang tidak relevan, serta menghilangkan duplikasi data yang dapat mengganggu analisis lebih lanjut.

Table 2. Contoh Data Hasil Cleaning

No	Comment	Cleaning
1	Iba 😂😂😂😂😂😂,,, liat rakyat mati2an cari sesuap nasi apa gak iba mata orang DPR	Iba liat rakyat matian cari sesuap nasi apa gak iba mata orang DPR

b. Case Folding

Proses ini bertujuan untuk merubah semua huruf dalam teks menjadi huruf kecil (*lowercase*) guna memastikan konsistensi dalam pemrosesan teks.

Table 3. Contoh Data Hasil Case Folding

No	Cleaning	Case Folding
1	Iba liat rakyat matian cari sesuap nasi apa gak iba mata orang DPR	iba liat rakyat matian cari sesuap nasi apa gak iba mata orang dpr

c. Normalisasi

Proses ini memperbaiki kata-kata slang atau kesalahan ketik dengan mengubahnya menjadi kata baku sesuai dengan KBBI (Kamus Besar Bahasa Indonesia).

Table 4. Contoh Data Hasil Normalisasi

No	Cleaning	Normalisasi
1	iba liat rakyat matian cari sesuap nasi apa gak iba mata orang dpr	iba lihat rakyat matian cari sesuap nasi apa tidak iba mata orang dpr

d. Tokenisasi

Tokenisasi adalah tahap di mana kalimat dipecah menjadi kata-kata berdasarkan spasi atau whitespace, yang memungkinkan analisis lebih mendalam terhadap setiap kata.

Table 5. Contoh Data Hasil Tokenisasi

No	Normalisasi	Tokenisasi
1	iba lihat rakyat matian cari sesuap nasi apa tidak iba mata orang dpr	['iba', 'lihat', 'rakyat', 'matian', 'cari', 'sesuap', 'nasi', 'apa', 'tidak', 'iba', 'mata', 'orang', 'dpr']

e. Stopword Removal

Jurnal Dinamika Informatika
Volume 15, No 1, Maret 2026
ISSN 1978-1660 : 29-43
ISSN *online* 2549-8517

Pada proses ini, kata-kata umum yang tidak memberikan informasi penting atau tidak relevan dengan analisis dihapus untuk meningkatkan kualitas data.

Table 6. Contoh Data Hasil Stopword Removal

No	Tokenisasi	Stopword Removal
1	['iba', 'lihat', 'rakyat', 'matian', 'cari', 'sesuap', 'nasi', 'apa', 'tidak', 'iba', 'mata', 'orang', 'dpr']	['iba', 'lihat', 'rakyat', 'matian', 'cari', 'sesuap', 'nasi', 'tidak', 'iba', 'mata', 'dpr']

f. Stemming

Proses ini mengubah kata menjadi kata dasarnya dengan menghapus imbuhan menggunakan algoritma Sastrawi, untuk menyederhanakan kata-kata dan memperkecil variasi bentuk kata.

Table 7. Contoh Data Hasil Stemming

No	Stopword Removal	Stemming
1	['iba', 'lihat', 'rakyat', 'matian', 'cari', 'sesuap', 'nasi', 'tidak', 'iba', 'mata', 'dpr']	iba lihat rakyat mati cari suap nasi tidak iba mata dpr

g. Klasifikasi Berbasis Lexicon dan IndoBERT

Pada tahap ini menggunakan algoritma hibrida yaitu gabungan metode berbasis leksikon dan model indoBERT, leksikon dilakukan dengan pemanfaatan kamus kata positif dan negatif untuk menghitung kecenderungan sentimen dan menentukan suatu komentar termasuk ke dalam sentimen positif, negatif dan netral.

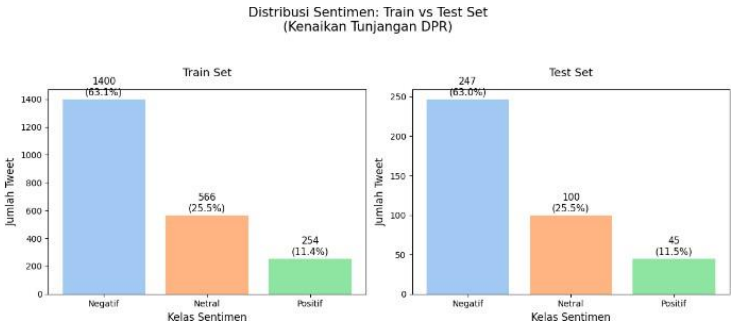
Selain metode leksikon penelitian ini juga menggunakan model pra-latih IndoBERT untuk klasifikasi sentimen secara kontekstual. IndoBERT dapat mendeteksi hubungan antar kata dalam suatu kalimat sehingga dapat memahami makna komentar lebih dalam. Hasil prediksi dari metode leksikon dan IndoBERT kemudian digabungkan menggunakan aturan tertentu untuk menghasilkan label sentimen akhir.

Table 8. Contoh Hasil Klasifikasi Sentimen

No	Komentar	Bobot Lexicon	Bobot IndoBERT	Label Sentimen
1	kompak stop bayar pajak	-0,8 (negatif)	0,98 (negatif)	negatif
2	ngeri bayar pajak kasih makan	-0,9 (negatif)	0,99 (negatif)	negatif
3	yuk gaskeun besok	+0,1 (positif)	0,88 (positif)	positif

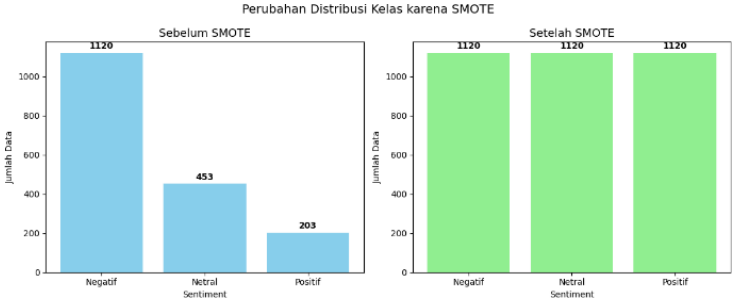
h. Pelatihan Model

Setelah data memperoleh label sentimen, proses selanjutnya yaitu pelatihan model klasifikasi. Data dibagi menjadi *train data* dan *tets data* dengan perbandingan 85:15.



Gambar 4. Split Data

Metode *Term Frequency-Inverse Document Frequency* (TF-IDF) digunakan sebagai metode representasi teks yang bertujuan untuk mengubah teks menjadi bentuk numerik. Dikarenakan adanya ketidakseimbangan jumlah data pada masing - masing kelas, maka dilakukan teknik SMOTE (*Synthetic Minority Over-sampling Technique*) yaitu mengenerate sampel data baru yang bersifat sintesis(buatan) untuk kelas minoritas sehingga meningkatkan representasi kelas minoritas.



Gambar 5. Hasil dari SMOTE

Dua algoritma yang digunakan pada penelitian ini yaitu *Support Vector Machine* (SVM) dan *Multinomial Naive Bayes* dibungkus menggunakan ImbPipeline untuk memastikan proses SMOTE hanya terjadi pada data *train* dan tidak membocorkan pada data *test*. Model pertama, *Support Vector Machine* (SVM) dengan kernel RBF, melalui proses penyetelan hiperparameter (GridSearchCV) dan menghasilkan konfigurasi optimal pada nilai $C = 10$ dan $\gamma = 1$, dengan skor validasi silang (CV score) mencapai 0,7297.

Jurnal Dinamika Informatika
Volume 15, No 1, Maret 2026
ISSN 1978-1660 : 29-43
ISSN *online* 2549-8517

```
✅ Best params (SVM): {'svm__C': 10, 'svm__gamma': 1}
✅ Best score (SVM): 0.7297297297297297

✅ Hasil evaluasi SVM (RBF):
Accuracy: 0.7657657657657657
              precision    recall  f1-score   support

     Negatif       0.80      0.94      0.86       280
      Netral       0.61      0.49      0.54       113
     Positif       0.88      0.43      0.58        51

    accuracy                           0.77       444
   macro avg       0.76      0.62      0.66       444
weighted avg       0.76      0.77      0.75       444
```

Gambar 6. Pelatihan Model SVM

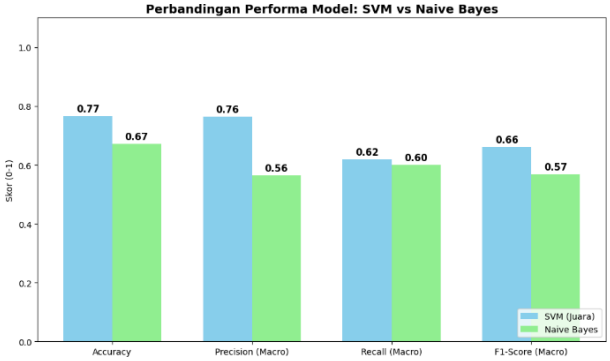
Sebagai pembanding, model *Multinomial Naive Bayes* juga dilatih dengan skema serupa dan menghasilkan parameter pemulusan (alpha) terbaik sebesar 0,5. Kedua model dengan kinerja terbaik tersebut disimpan dalam format .pkl.

```
✅ Melatih model Multinomial Naive Bayes dengan GridSearch (versi cepat)...
Fitting 2 folds for each of 2 candidates, totalling 4 fits

✅ Best params (NB): {'nb__alpha': 0.5}
Accuracy: 0.6711711711711712
              precision    recall  f1-score   support

     Negatif       0.81      0.80      0.81       280
      Netral       0.51      0.37      0.43       113
     Positif       0.38      0.63      0.47        51

    accuracy                           0.67       444
   macro avg       0.56      0.60      0.57       444
weighted avg       0.68      0.67      0.67       444
```

Gambar 7. Pelatihan Model Naive Bayes

i. Evaluasi dan Pembahasan Hasil

Evaluasi model dilakukan untuk mengetahui tingkat keberhasilan model dalam mengklasifikasikan sentimen komentar. Metode evaluasi yang digunakan confusion matrix dengan matrix akurasi, precision, recall, dan F1 score.

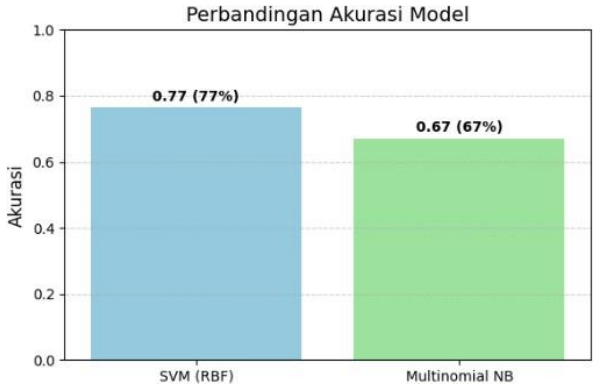


Gambar 8. Perbandingan Performa Model

Hasil pengujian menunjukkan bahwa model *Support Vector Machine* (SVM) memiliki performa yang lebih baik dibandingkan *Multinomial Naive Bayes*. Model SVM mampu mencapai tingkat akurasi sebesar 76,6%, sementara *Multinomial Naive Bayes* mendapatkan akurasi sebesar 67,1%.

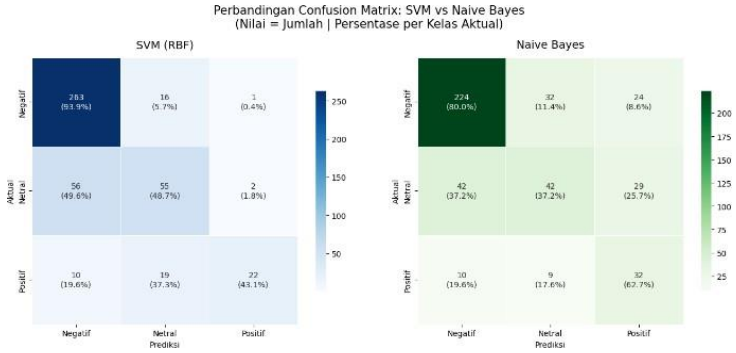
Hasil penelitian ini konsisten dengan penelitian oleh Permana et al. (2025) yang membandingkan performa algoritma Support Vector Machine

dan Naive Bayes dalam analisis sentimen komentar YouTube. Penelitian yang dilakukan pada konteks komentar YouTube terkait industri esports di Indonesia menunjukkan bahwa SVM memiliki performa yang sedikit lebih unggul dibandingkan Naive Bayes dalam mengklasifikasikan sentimen negatif dan netral, meskipun perbedaan nilai akurasi dan F1-score tidak terlalu signifikan. Temuan tersebut mendukung penggunaan algoritma SVM pada data komentar YouTube berbahasa Indonesia yang memiliki karakteristik bahasa tidak terstruktur dan beragam [11].



Gambar 9. Perbandingan Akurasi Model

Hasil confusion matrix menunjukkan bahwa model memiliki performa yang baik dalam mengklasifikasikan sentimen negatif, namun masih mengalami kesulitan dalam membedakan sentimen positif dan netral.



Gambar 10. Confusion Matrix

Temuan ini mengindikasikan bahwa ekspresi sentimen negatif pada komentar cenderung memiliki pola bahasa yang lebih konsisten, sedangkan sentimen positif dan netral memiliki variasi bahasa yang lebih beragam sehingga lebih sulit diklasifikasikan secara akurat.

j. Keterbatasan dan Implikasi

Meskipun penelitian ini memperlihatkan hasil yang cukup baik, terdapat beberapa keterbatasan yang perlu diperhatikan. Pertama, proses pelabelan sentimen dilakukan secara otomatis menggunakan pendekatan hibrida tanpa validasi annotator manusia, sehingga potensi bias dalam pelabelan masih dimungkinkan. Kedua, model IndoBERT yang digunakan belum melalui proses fine-tuning secara khusus pada dataset penelitian.

Jurnal Dinamika Informatika
Volume 15, No 1, Maret 2026
ISSN 1978-1660 : 29-43
ISSN *online* 2549-8517

Implikasi dari penelitian ini menunjukkan bahwa pendekatan hibrida lexicon dan IndoBERT dapat dimanfaatkan sebagai solusi awal dalam analisis sentimen media sosial. Namun, untuk memperoleh hasil yang lebih akurat, penelitian selanjutnya disarankan untuk menggunakan data berlabel manual serta melakukan fine-tuning model bahasa agar mampu menangkap konteks sentimen secara lebih optimal.

KESIMPULAN

Berdasarkan hasil penelitian dan pembahasan yang telah dilakukan, dapat ditarik kesimpulan sebagai berikut:

1. Pendekatan pelabelan sentimen otomatis yang mengombinasikan metode berbasis leksikon dan model IndoBERT dapat digunakan untuk menghasilkan dataset berlabel pada komentar YouTube yang sebelumnya tidak memiliki label, sehingga mendukung proses pelatihan model klasifikasi sentimen.
2. Hasil evaluasi menunjukkan bahwa algoritma Support Vector Machine (SVM) memiliki performa yang lebih baik dibandingkan Multinomial Naive Bayes dalam mengklasifikasikan sentimen komentar masyarakat terhadap isu yang dianalisis.
3. Meskipun demikian, performa kedua model klasifikasi tersebut masih belum optimal dalam membedakan kelas sentimen positif dan netral secara akurat.
4. Keterbatasan performa model dipengaruhi oleh karakteristik data media sosial yang memiliki variasi bahasa tidak baku, penggunaan ekspresi informal, serta adanya makna implisit yang sulit dipahami oleh model klasifikasi.
5. Penerapan teknik penyeimbangan data menggunakan SMOTE dapat meningkatkan representasi kelas minoritas dalam dataset, namun belum sepenuhnya mampu mengatasi kompleksitas semantik yang terdapat pada komentar media sosial.

SARAN

Untuk meningkatkan kualitas hasil analisis pada penelitian selanjutnya, disarankan agar proses pelabelan sentimen tidak hanya dilakukan secara otomatis, tetapi juga dilengkapi dengan validasi manual oleh manusia guna memastikan keakuratan label data. Selain itu, model IndoBERT yang digunakan dapat ditingkatkan kinerjanya melalui proses *fine-tuning* dengan dataset yang lebih sesuai dengan karakteristik bahasa media sosial berbahasa Indonesia.

Selanjutnya, penggunaan teknik *data augmentation* disarankan untuk menambah variasi data dan membantu mengatasi ketidakseimbangan jumlah data pada setiap kelas sentimen. Penelitian selanjutnya juga dapat mengembangkan hasil

analisis ini ke dalam bentuk aplikasi atau dashboard analitik, sehingga hasil penelitian tidak hanya bersifat teoritis, tetapi juga dapat dimanfaatkan secara praktis untuk memantau opini publik.

DAFTAR PUSTAKA

[1] A. M. A. Ausat, 2023, The Role of Social Media in Shaping Public Opinion and Its In- fluence on Economic Decisions, No.1, Vol.1, 35–44, doi: 10.61100/tacit.v1i1.37.

[2] M. N. Alkhudari, O. J. Abduljabbar, A. Mohammed, and A. Manaseer, 2024, The Role of Social Media in Shaping Public Opinion Among Jordanian University Students, *J. Infrastructure, Policy Dev.*, No.8, Vol.8, 1–25, doi: 10.24294/jipd.v8i8.5489.

[3] S. S. Shah, 2024, The Role of Social Media in Shaping Public Opinion on Environmental Issues, *Prem. J. Environ. Sci.*, Vol.1, doi: 10.70389/PJES.100002.

[4] H. Ju, H. Lee, J. Choi, and E. Kang, 2025, The Necessity of Regulating Drinking Scenes on Social Media Platforms Focusing on YouTube Sulbang Videos : Public Opinion From Surveys and YouTube Content Analysis, *JMIR Form. Res.*, Vol.9, doi: 10.2196/65162.

[5] T. Hidayat, A. Bahtiar, and Kaslani, 2025, Classification of Youtube User sentiment on 5G Technology Videos with Naïve Bayes, *J. Artif. Intell. Eng. Appl.*, No.3, Vol.4, 1706–1711, doi: 10.59934/jaiea.v4i3.992.

[6] V. B. Lestari and C. A. Hutagalung, 2025, Evaluation of TF-IDF Extraction Techniques in Sentiment Analysis of Indonesian-Language Marketplaces Using SVM , Logistic Regression , and Naive Bayes, *J-KOMA J. Comput. Sci. Appl.*, No.1. Vol.8, 36–44, doi: 10.21009/j-koma.v8i1.05.

[7] Y. Asri, D. Kuswardani, W. N. Suliyanti, Y. O. Manullang, and A. R. Ansyari, 2025, Sentiment Analysis Based on Indonesian Language Lexicon and IndoBERT on User Reviews PLN Mobile Application, *Indones. J. Electr. Eng. Comput. Sci.*, No.1, Vol.38, 677–688, doi: 10.11591/ijeecs.v38.i1.pp677-688.

[8] N. V Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, 2002, SMOTE : Synthetic Minority Over-sampling Technique, *J. Artif. Intell. Res.*, Vol.16, 321–357, doi: 10.1613/jair.953.

[9] S. F. Huwaida, R. Kusumawati, and B. Isnaini, 2024, Analisis Sentimen Komentar Youtube terhadap Pemindahan Ibu Kota Negara Menggunakan Metode Naïve Bayes, *JAMBURA J. INFORMATICS*, No.1, Vol.6, 26–39, doi: 10.37905/jji.v6i1.24718.

[10] D. M. Christopher, P. Raghavan, and H. Schütze, 2009, *An Introduction to Information Retrieval*. Cambridge.

[11] T. D. Permana, Y. B. Pratama, Z. Wahyuzi, E. Altiarika, and A. Pramudyantoro, 2025, Perbandingan Performa Algoritma Naive Bayes dan SVM untuk Analisis Sentimen Komentar YouTube terhadap Industri Esports

di Indonesia, *J. Ilm. Nusant.*, No.6, Vol.2, 1391–1399, doi: 10.61722/jinu.v2i6.6753.