Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 4, Juni 2026
e-ISSN 2685-998X
DOI 10.30865/json.v7i4.9632

Perbandingan SVM dan Logistic Regression Berbasis SMOTE pada Analisis Sentimen Menteri Keuangan di YouTube

Sirozul Huda*, Afril Efan Pajri, Sahri
Teknik Informatika, Fakultas Sains dan Teknologi, Universitas Nahdlatul Ulama Sunan Giri, Bojonegoro, Indonesia
Email: [1,*]sirozulhuda18@gmail.com, [2]afril@unugiri.ac.id, [3]sahriunugiri@gmail.com
Email Penulis Korespondensi: sirozulhuda18@gmail.com*
Submitted: **01/03/2026**; Accepted: **28/03/2026**; Published: **30/06/2026**

Abstrak- YouTube telah berkembang menjadi salah satu ruang utama bagi masyarakat untuk menyampaikan opini terhadap isu publik, termasuk respons terhadap figur pejabat pemerintah. Penelitian ini bertujuan untuk mengevaluasi dan memetakan persepsi publik terhadap Menteri Keuangan Republik Indonesia melalui analisis sentimen berbasis machine learning, dengan mengkomparasikan kinerja algoritma *Support Vector Machine* dan *Logistic Regression*. Data berjumlah 4.003 data komentar dikumpulkan dari komentar berbagai kanal YouTube berdasarkan pencarian topik terkait, kemudian melalui tahapan pre-processing dan pelabelan otomatis menggunakan Indonesian Sentiment Lexicon (InSet). Meskipun pelabelan otomatis berbasis leksikon ini efisien, pendekatan ini memiliki keterbatasan untuk mendeteksi sarkasme dan menangkap konteks kalimat yang kompleks. Setelah fitur diekstraksi menggunakan pembobotan TF-IDF, teknik *Synthetic Minority Over-sampling Technique* (SMOTE) diterapkan untuk mengatasi masalah ketidakseimbangan label pada data latih. Pengujian model membuktikan bahwa pendekatan yang diusulkan berhasil melakukan klasifikasi dengan sangat baik. Hasil evaluasi menunjukkan bahwa *Support Vector Machine* memberikan tingkat akurasi tertinggi sebesar 90%, sedikit mengungguli *Logistic Regression* yang mencatatkan akurasi sebesar 89%. Kontribusi ilmiah dalam studi ini menegaskan bahwa algoritma klasifikasi berbasis SMOTE efektif dalam menangani masalah ketimpangan data, dengan menunjukkan peningkatan performa model dibandingkan penelitian terdahulu yang serupa.

Kata Kunci: Analisis Sentimen; Menteri Keuangan; Support Vector Machine; Logistic Regression; SMOTE

Abstract- YouTube has evolved into one of the primary platforms for the public to express opinions on public issues, including reactions to government officials. This study aims to evaluate and map public perceptions of the Minister of Finance of the Republic of Indonesia through machine learning-based sentiment analysis, comparing the performance of the Support Vector Machine and Logistic Regression algorithms. A total of 4,003 comment data points were collected from comments across various YouTube channels based on searches for related topics, followed by pre-processing and automatic labeling using the Indonesian Sentiment Lexicon (InSet). Although this lexicon-based automatic labeling is efficient, this approach has limitations in detecting sarcasm and capturing complex sentence contexts. After features were extracted using TF-IDF weighting, the Synthetic Minority Over-sampling Technique (SMOTE) was applied to address the issue of label imbalance in the training data. Model testing demonstrated that the proposed approach successfully performed classification very well. Evaluation results show that the Support Vector Machine achieved the highest accuracy rate of 90%, slightly outperforming Logistic Regression, which recorded an accuracy of 89%. The scientific contribution of this study confirms that SMOTE-based classification algorithms are effective in addressing data imbalance issues, demonstrating improved model performance compared to similar previous studies.

Keywords: Sentiment Analysis; Minister of Finance; Support Vector Machine; Logistic Regression; SMOTE

1. PENDAHULUAN

Dinamika politik dan ekonomi nasional mengalami transisi signifikan menyusul pelantikan Purbaya Yudhi Sadewa sebagai Menteri Keuangan Republik Indonesia pada 9 September 2025, menggantikan posisi yang sebelumnya diemban oleh Sri Mulyani Indrawati [1]. Pelantikan tersebut turut menjadi perhatian publik, terutama karena rekam jejak Purbaya Yudhi Sadewa yang sebelumnya menjabat sebagai Ketua Dewan Komisioner Lembaga Penjamin Simpanan (LPS), serta pandangannya yang kritis terhadap kebijakan perpajakan dan penerimaan negara [2]. Beberapa media dalam negeri menyampaikan bahwasanya substitusi Menteri Keuangan tersebut memicu respons impulsif dari pasar keuangan, di mana menurunnya Indeks Harga Saham Gabungan (IHSG) dan depresiasi mata uang rupiah [3].

Diskusi mengenai kebijakan dan isu publik tidak hanya terjadi di ruang formal, tetapi juga diperluas pada platform media sosial. Di antara berbagai platform yang tersedia, YouTube menjadi salah satu media yang paling banyak digunakan [4]. Menurut Data Reportal, Indonesia memiliki kurang lebih 143 juta pengguna aktif [5]. Platform ini telah bertransformasi dari sekadar media hiburan visual menjadi ruang dialektika bagi publik secara terbuka. Karakteristik konten YouTube yang memiliki durasi tayangan panjang memberikan ruang bagi audiens untuk memahami konteks suatu isu secara lebih utuh. Kondisi ini secara langsung mendorong munculnya interaksi di kolom komentar yang cenderung lebih tajam dan argumentatif terhadap opini masyarakat arus bawah terkait sosok dan juga kinerja Menteri Keuangan Purbaya Yudhi Sadewa.

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 4, Juni 2026
e-ISSN 2685-998X
DOI 10.30865/json.v7i4.9632

Dalam konteks dinamika sosial dan politik yang berlangsung, pemetaan persepsi publik menjadi salah satu pendekatan yang penting untuk menelusuri bagaimana pandangan masyarakat terbentuk melalui berbagai interaksi di media sosial [6]. Namun, tantangan muncul karena media sosial menghasilkan data dalam jumlah yang sangat besar, sehingga proses pemetaan secara manual menjadi kurang efisien. Kondisi tersebut menjadikan analisis sentimen berbasis *machine learning* semakin diperlukan untuk melakukan pemantauan opini publik secara otomatis [7]. Secara umum, analisis sentimen dapat dipahami sebagai metode komputasi yang digunakan untuk mengekstraksi dan menafsirkan pandangan, respons, maupun kesan masyarakat terhadap berbagai isu, layanan, atau produk [8]. Secara teknis, analisis sentimen merupakan metode guna mengidentifikasi dan mengekstrak informasi subjektif dari data teks melaui pemanfaatan teknik penambangan teks (*text mining*) dan pemrosesan bahasa alami (*Natural Language Processing*) untuk mengklasifikasikan data ke dalam kategori yang spesifik, seperti sentimen positif atau negatif [9].

Dalam ranah klasifikasi teks komputasional, *Support Vector Machine* (SVM) dan *Logistic Regression* merupakan dua algoritma *machine learning* yang masif digunakan. Kedua algoritma tersebut dipilih karena memiliki karakteristik yang dianggap sesuai untuk pengolahan data teks dan telah banyak digunakan dalam berbagai studi serupa. Pada model SVM, proses klasifikasi dilakukan dengan mencari garis batas pemisah (*hyperplane*) yang paling tepat di ruang fitur berdimensi tinggi untuk membedakan komentar bernada positif dan negatif dengan jarak pemisahan yang maksimal[10]. Di sisi lain, *Logistic Regression* memperkirakan kemungkinan suatu data berada pada kelas tertentu berdasarkan keterkaitan antara variabel target dan sejumlah fitur yang digunakan. Pendekatan ini memungkinkan interpretasi yang lebih jelas terhadap peran setiap fitur dalam menentukan hasil klasifikasi [11].

Hadi dan Sugiarto [12] melakukan penelitian analisis sentimen publik pada media sosial X dengan mengangkat isu pembangunan Ibu Kota Nusantara. Dalam kajian tersebut, beberapa algoritma klasifikasi dibandingkan untuk melihat perbedaan kinerjanya. Hasil analisis menunjukkan bahwa *Support Vector Machine* menghasilkan kinerja terbaik dengan tingkat akurasi sebesar 80%, sementara *Logistic Regression* dan *Naive Bayes* mencatatkan akurasi yang sedikit lebih rendah, yaitu sebesar 79%. Penelitian lain yang dilakukan oleh Maulana dkk. [13] menelaah opini pengguna terhadap aplikasi Gojek melalui ulasan di Google Play Store. Hasil analisis menunjukkan bahwa *Logistic Regression* menghasilkan tingkat akurasi tertinggi, yaitu sebesar 82,45%. Sebaliknya, *K-Nearest Neighbor* menunjukkan kinerja paling rendah dengan akurasi sebesar 52,28%.

Ainunnisa dan Sulastri [14] mengkaji sentimen pengguna pada ulasan aplikasi hiburan dengan membandingkan sejumlah algoritma klasifikasi. Temuan penelitian tersebut menunjukkan bahwa *Logistic Regression* menghasilkan kinerja paling optimal dengan tingkat akurasi sebesar 84 %, diikuti oleh *Support Vector Machine* dengan akurasi sebesar 82%. Sementara itu, Dwinnie dan Novita [15] menelaah respons publik terhadap debat calon presiden dan wakil presiden melalui media sosial X. Hasil kajian tersebut memperlihatkan bahwa *Support Vector Machine* menghasilkan tingkat akurasi sebesar 78%, sedangkan *Logistic Regression* menunjukkan kinerja yang relatif lebih baik dengan akurasi mencapai 79%. Lebih lanjut, Prasetyo dkk. [16] melakukan kajian dengan melakukan komparasi algoritma *Logistic Regression*, SVM, dan Random Forest pada ulasan aplikasi dompet digital Gopay. Hasil menunjukkan akurasi tertinggi diperoleh *Logistic Regression* dengan 88,16%, mengungguli SVM 87,5% dan *Random Forest* 79,33%.

Berbagai kajian literatur telah membuktikan kehandalan SVM dan *Logistic Regression*, terdapat satu masalah fundamental yang sering muncul dalam dataset media sosial, yakni fenomena ketidakseimbangan kelas (*imbalanced data*). Dalam konteks komentar YouTube yang menyoroti pejabat publik, kecenderungan psikologis warganet untuk melontarkan kritik sering kali menghasilkan dominasi sentimen negatif yang sangat timpang dibandingkan sentimen positif. Jika dataset yang tidak seimbang ini langsung diproses, model SVM maupun *Logistic Regression* akan cenderung mengabaikan pola dari kelas minoritas dan menghasilkan prediksi klasifikasi yang bias (*overfitting*) terhadap kelas mayoritas [17]. Guna mengatasi anomali distribusi tersebut, penelitian ini mengusulkan penerapan *Synthetic Minority Over-sampling Technique* (SMOTE). Teknik prapemrosesan ini beroperasi dengan menyintesis data buatan pada kelas minoritas berdasarkan kedekatan ruang fitur (*k-nearest neighbors*), sehingga proporsi antar kelas sentimen menjadi seimbang sebelum model dilatih. Integrasi teknik SMOTE diproyeksikan mampu menekan bias algoritma dan mengekstraksi performa klasifikasi yang sesungguhnya [18].

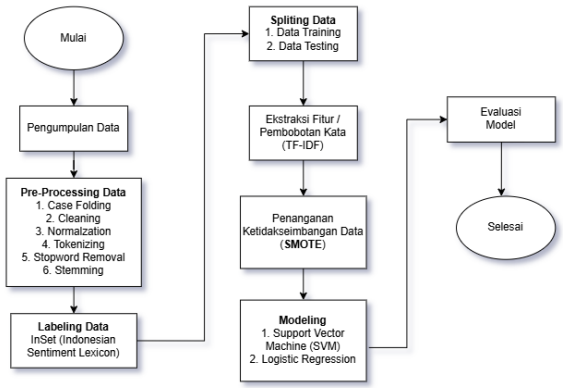
Berdasarkan pemetaan literatur di atas, terlihat adanya celah penelitian (*research gap*) yang signifikan. Mayoritas studi sentimen publik masih berkutat pada ulasan aplikasi digital atau kebijakan umum berskala luas, dan lebih banyak mengeksploitasi data dari platform X. Riset yang secara spesifik membedah sentimen publik terhadap figur pucuk pimpinan kementerian, memanfaatkan karakteristik komentar di YouTube, serta mengevaluasi komparasi algoritma melalui penanganan *imbalanced data* menggunakan SMOTE, masih terbatas. Oleh karena itu, penelitian ini bertujuan untuk mengkomparasikan kinerja *Support Vector Machine* dan *Logistic Regression* yang dioptimasi menggunakan SMOTE dalam mengklasifikasikan polaritas sentimen publik terhadap Menteri Keuangan Republik Indonesia. Hasil evaluasi model ini diharapkan mampu memberikan kontribusi metodologis dalam skenario data teks tidak seimbang, sekaligus menyajikan wawasan empiris bagi pemerintah mengenai tingkat penerimaan masyarakat di ruang digital.

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 4, Juni 2026
e-ISSN 2685-998X
DOI 10.30865/json.v7i4.9632

2. METODOLOGI PENELITIAN

2.1 Tahapan Penelitian

Penelitian ini dibagi secara sistematis menjadi beberapa tahapan berurutan. Pengumpulan data, *pre-processing* data, pelabelan data dengan metode *lexicon based*, dan pembagian data merupakan langkah-langkah awal dalam proses ini. Metode pembobotan TF-IDF digunakan untuk ekstraksi fitur, teknik SMOTE digunakan untuk mengatasi ketidakseimbangan kelas data, dan algoritma *Support Vector Machine* (SVM) dan *Logistic Regression* digunakan untuk pemodelan. Fase evaluasi model digunakan untuk mengukur dan membandingkan performa kedua model. Gambar 1 di bawah menampilkan seluruh tahapan pada penelitian.



Gambar 1. Tahapan Penelitian

2.2 Pengumpulan Data

Pengumpulan data dilaukan secara otomatis dengan menggunakan bahasa pemrograman *Python* yang diintegrasikan dengan *YouTube Data API v3*. API ini digunakan untuk membuka akses terstruktur terhadap metadata konten dan repositori komentar publik. Proses ekstraksi difokuskan pada video-video yang relevan dengan diskursus publik mengenai Menteri Keuangan Republik Indonesia pada fase krusial awal masa jabatannya, yakni dalam rentang tanggal 8 September 2025 hingga 31 Oktober 2025.

Pencarian target data dilakukan secara spesifik menggunakan query "Menteri Keuangan Purbaya Yudhi", dengan batasan parameter penarikan diatur pada maksimal 20 video teratas. Berdasarkan identifikasi Video ID dari hasil pencarian tersebut, sistem kemudian melakukan *scraping* terhadap seluruh komentar yang tersedia pada masing-masing tayangan. Tahapan ini berhasil mengumpulkan total sebanyak 4.003 baris data komentar mentah (*raw data*) yang selanjutnya akan diproses pada tahap *pre-processing*.

2.3 *Pre-processing Data*

Pre-processing dilakukan sebagai tahap krusial awal untuk membersihkan data teks mentah yang diperoleh melalui proses scraping, sehingga korpus data tersebut siap digunakan pada tahapan komputasi selanjutnya [19]. Tujuan utama dari tahapan ini adalah untuk meminimalkan data yang bersifat noisy sekaligus meningkatkan kualitas fitur yang akan diproses oleh model klasifikasi. Rangkaian pembersihan ini diimplementasikan menggunakan skrip Python melalui enam tahapan berurutan. Proses diawali dengan *case folding* untuk menyeragamkan seluruh karakter menjadi huruf kecil, dilanjutkan dengan tahap *cleaning* yang memanfaatkan operasi *Regular Expression* (*Regex*) guna mengeliminasi elemen non-tekstual seperti *URL*, *hashtag*, tautan *mention* , tanda baca, angka, dan emoji [20].

Mengingat data bersumber dari kolom komentar YouTube yang sering menggunakan bahasa tidak baku, tahap normalisasi diaplikasikan untuk menstandarisasi *slang words* dan singkatan menjadi kata baku. Proses pemetaan dan perbaikan kata ini menggunakan kamus translasi khusus (*slang dictionary*) yang memuat kompilasi 1.599 pasang kata tidak baku beserta kata bakunya [21]. Teks yang telah terstandardisasi kemudian dipecah menjadi satuan kata tunggal melalui tahapan tokenizing [22]. Selanjutnya, kata-kata hubung yang memiliki frekuensi kemunculan tinggi namun tidak memberikan kontribusi informatif terhadap makna sentimen dieliminasi melalui tahap *stopword removal*. Proses ini menggunakan leksikon bawaan bahasa Indonesia dari pustaka *Natural Language Toolkit* (NLTK), yang dimodifikasi secara kustom. Modifikasi dilakukan dengan menambahkan kata-kata pengganggu khas media sosial, sekaligus secara krusial mengecualikan kata-kata negasi (seperti "tidak", "kurang", "jangan") dari daftar penghapusan guna memastikan konteks polaritas sentimen tidak hilang. Rangkaian

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 4, Juni 2026
e-ISSN 2685-998X
DOI 10.30865/json.v7i4.9632

pre-*processing* ini diakhiri dengan tahap *stemming* menggunakan algoritma dari pustaka Sastrawi untuk mereduksi dan menghilangkan seluruh imbuhan sehingga setiap kata kembali ke bentuk dasarnya [23].

2.4 Pelabelan Data

Pelabelan data dilakukan secara otomatis menggunakan metode *Lexicon-based* melalui pendekatan kamus *Indonesian Sentiment Lexicon* (InSet). Setiap kata dalam kamus InSet memiliki bobot polaritas yang telah ditetapkan, berkisar antara -5 hingga +5. Bobot sentimen untuk sebuah komentar dihitung dengan mengakumulasi nilai bobot dari seluruh kata yang menyusun kalimat komentar tersebut. Dalam penelitian ini, klasifikasi dikonfigurasi secara biner ke dalam dua kelas sentimen utama. Apabila hasil akumulasi perhitungan menunjukkan skor akhir lebih besar dari nol (> 0), maka teks secara otomatis diklasifikasikan ke dalam kelas sentimen positif. Sebaliknya, jika hasil perhitungan bernilai kurang dari atau sama dengan nol (≤ 0), teks tersebut secara tegas diklasifikasikan sebagai kelas sentimen negatif. Penggabungan nilai persis nol ke dalam batas kelas negatif ini diterapkan untuk mendikotomikan seluruh sampel data tanpa menyisakan kelas netral, sehingga memfokuskan model pada deteksi narasi yang murni mendukung (positif) dengan narasi yang mengkritik atau tidak memiliki sentimen dukungan sama sekali (negatif).

Meskipun metode ini efisien secara waktu komputasi untuk memberikan label pada ribuan data teks, pendekatan leksikon memiliki titik lemah bawaan yang menjadi batasan (*limitation*) dalam penelitian ini. Karena leksikon bekerja dengan cara memetakan bobot kata secara individual, metode ini rentan mengalami *misclassification* (salah label) saat menghadapi data dengan bahasa yang kompleks. Berdasarkan evaluasi yang dilakukan oleh pencipta kamus InSet itu sendiri, pendekatan leksikon terbukti memiliki kelemahan mendasar karena tidak mampu mengenali konteks kata secara utuh dan gagal mendeteksi keberadaan kalimat sarkasme [24]. Keterbatasan dalam menangkap makna semantik tingkat tinggi ini, khususnya pada komentar bermuatan politik di YouTube yang sarat akan sindiran, disadari sebagai salah satu faktor yang memengaruhi validitas murni dari pelabelan kelas otomatis sebelum didistribusikan sebagai data latih algoritma klasifikasi.

2.5 Pembagian Data

Setelah pelabelan data, langkah selanjutnya adalah membagi dataset menjadi dua himpunan bagian yang saling lepas, yaitu data latih (*training data*) dan data uji (*testing data*). Proses pemisahan ini dilakukan secara acak (*random sampling*) guna memastikan distribusi data yang representatif pada kedua bagian [25].

Dalam penelitian ini, pembagian data dilakukan dengan proporsi rasio 80:20 dari total dataset bersih yang berjumlah 3.896 baris. Sebanyak 80% dari total data, yaitu sejumlah 3.116 baris dialokasikan sebagai data latih, yang berfungsi sebagai materi pembelajaran bagi algoritma Support Vector Machine dan Logistic Regression untuk mengenali pola sentimen. Sementara itu, sisa 20% data, yaitu sejumlah 779 baris dialokasikan sebagai data uji.

2.6 Ekstraksi Fitur (TF-IDF)

Term Frequency–Inverse Document Frequency atau TF-IDF digunakan untuk merepresentasikan bobot kata dalam dokumen teks berdasarkan tingkat kemunculannya [26]. Pendekatan ini mengombinasikan informasi frekuensi kata pada satu dokumen dengan tingkat kemunculan kata tersebut di seluruh kumpulan dokumen. Nilai *Term Frequency* menggambarkan intensitas kemunculan suatu kata dalam sebuah dokumen, yang dihitung menggunakan persamaan 1.

$$TF(t,d) = \frac{f_{t,d}}{\sum_k f_{k,d}} \tag{1}$$

Keterangan:

a. $TF(t,d)$ = nilai frekuensi kata dalam satu dokumen.
b. $f_{t,d}$= jumlah kemunculan term t dalam dokumen d.
c. $\sum_k f_{k,d}$ = total frekuensi seluruh kata pada dokumen d.

Inverse Document Frequency digunakan untuk menggambarkan tingkat kekhasan suatu kata dalam keseluruhan korpus. Kata yang kemunculannya relatif terbatas pada sejumlah kecil dokumen akan memperoleh bobot yang lebih besar [27]. Perhitungan nilai IDF ditunjukkan pada persamaan 2.

$$IDF(t) = \log\left(\frac{N}{dF_t}\right) \tag{2}$$

Keterangan:

a. $IDF(t)$ = nilai bobot kelangkaan kata dalam seluruh dokumen.
b. N= jumlah keseluruhan dokumen dalam korpus.
c. dF_t= jumlah dokumen yang di dalamnya terdapat term (t).

$$TF - IDF(t,d) = TF(t,d) \times IDF(t) \tag{3}$$

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 4, Juni 2026
e-ISSN 2685-998X
DOI 10.30865/json.v7i4.9632

Keterangan:

$TF - IDF(t, d)$ = hasil penggabungan antara *TF dan IDF*.

Nilai tersebut dimanfaatkan untuk merepresentasikan tingkat kepentingan sebuah kata dalam dokumen dengan mempertimbangkan sebarannya pada seluruh kumpulan data. Pada implementasinya, transformasi teks menjadi matriks bobot numerik ini dieksekusi memanfaatkan modul *TfidfVectorizer* dari *Scikit-Learn* dengan pengaturan parameter *max-features* (5000) dan *n-gram* sebesar (1,2).

2.7 *Handling Imbalance Data (SMOTE)*

Dalam dataset analisis sentimen, kerap ditemukan fenomena ketidakseimbangan kelas (*class imbalance*) yang dapat menyebabkan model bias terhadap kelas mayoritas dan gagal mengenali pola kelas minoritas [17]. Guna mengatasi masalah tersebut, penelitian ini mengimplementasikan metode *Synthetic Minority Over-sampling Technique* (SMOTE). Berbeda dengan random over-sampling konvensional yang sekadar menduplikasi data dan rentan memicu *overfitting*, SMOTE bekerja secara dinamis dengan membangkitkan data sintetis baru.

Mekanisme SMOTE dilakukan dengan mengidentifikasi tetangga terdekat (*k-nearest neighbors*) pada sampel kelas minoritas, lalu menginterpolasi titik data sintetis di sepanjang garis yang menghubungkan sampel target dengan tetangganya [18]. Secara matematis, formulasi pembentukan data baru ditunjukkan pada Persamaan 4.

$$x_{new} = x_i + (x_{zi} - x_i) \times \mathrm{R} \tag{4}$$

Keterangan:

a. x_{new} : vektor fitur dari sampel minoritas asli.
b. x_i : vektor fitur dari salah satu tetangga terdekat.
c. x_{zi}: vektor fitur dari tetangga terdekat (*k-nearest neighbor*) yang terpilih.
d. R : bilangan acak antara 0 dan 1.

Pada implementasinya, teknik over-sampling ini dieksekusi menggunakan pustaka *Imbalanced-Learn* dengan pengaturan *k_neighbors* = 5 dan *random_state* = 42 guna memastikan replikabilitas eksperimen. Secara krusial, proses sintesis data ini diaplikasikan secara eksklusif hanya pada himpunan data latih. Langkah ini wajib dilakukan untuk mencegah kebocoran data (*data leakage*) sekaligus memastikan data uji tetap merepresentasikan kondisi ketimpangan data.

2.8 *Support Vector Machine*

Support Vector Machine (SVM) merupakan algoritma klasifikasi yang berupaya menemukan batas pemisah terbaik (*hyperplane*) untuk membedakan data berdasarkan kategorinya [28]. Pada ruang fitur berdimensi tinggi yang dihasilkan dari ekstraksi TF-IDF, SVM bekerja dengan membentuk pemisah linier yang secara optimal memisahkan titik-titik data ke dalam dua kutub sentimen (klasifikasi biner). Dalam klasifikasi teks linier ini, penentuan kelas sentimen dari sebuah korpus uji yang baru dieksekusi menggunakan fungsi keputusan matematis yang diformulasikan pada Persamaan 5.

$$f(x) = \mathrm{sign}(w \cdot x + b) \tag{5}$$

Keterangan:

a. $f(x)$: fungsi keputusan yang menghasilkan prediksi kelas akhir (misalnya +1 untuk kelas positif, dan -1 untuk kelas negatif).
b. w : vektor bobot (*weight vector*) yang merepresentasikan nilai kepentingan setiap fitur kata terhadap penentuan kelas.
c. x: vektor fitur input (nilai pembobotan TF-IDF dari teks yang akan diklasifikasikan).
d. b : nilai bias yang berfungsi untuk menggeser posisi *hyperplane* agar pemisahan kelas menjadi lebih presisi.
e. sign: fungsi tanda yang akan memetakan total hasil penjumlahan matematis menjadi keputusan biner.

Melalui persamaan tersebut, setiap kata di dalam kalimat akan dikalikan dengan bobotnya masing-masing, kemudian diakumulasikan dan ditambah dengan nilai bias untuk mendapatkan skor akhir. Skor margin inilah yang menentukan di sisi *hyperplane* mana suatu data tersebut diproyeksikan. Pada tahap implementasi eksperimen, algoritma ini dibangun memanfaatkan pustaka *Scikit-Learn*, dengan konfigurasi parameter *kernel = 'linear'*, nilai penalti margin = 1, serta *random_state* = 42 guna menjaga konsistensi output model.

2.9 *Logistic Regression*

Logistic Regression merupakan salah satu metode klasifikasi yang umum diterapkan dalam analisis data untuk memodelkan peluang suatu data berada pada kelas tertentu [29]. Metode ini memanfaatkan hubungan antara variabel target dan sejumlah fitur, serta bekerja secara optimal ketika keterkaitan antarvariabel bersifat linear. Dua persamaan utama digunakan untuk menghitung *logistic regression* secara matematis, yang ditunjukkan pada persamaan 6 dan 7 berikut.

$$W = b + w_1X_1 + w_2X_2 + w_3X_3 + \cdots + w_nX_n \tag{6}$$

Keterangan:

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 4, Juni 2026
e-ISSN 2685-998X
DOI 10.30865/json.v7i4.9632

a. W = nilai bobot akhir.
b. b = nilai bias.
c. $w_1\ w_2 \ldots, w_n$ = bobot atau koefisien yang diberikan pada masing-masing fitur
d. $X_1, X_2, X_3, \ldots, X_n$ = nilai fitur atau atribut pada sebuah sampel

$$P = \frac{1}{1 + e^{-W}} \tag{7}$$

Keterangan:
a. P = nilai probabilitas
b. e = bilangan eksponensial (≈2.718).

Fungsi sigmoid pada persamaan 7 mengubah nilai linear W menjadi probabilitas antara 0 dan 1. Semakin besar nilai W, semakin tinggi kemungkinan data akan diklasifikasikan ke dalam kelas positif. Semakin besar nilai positif W yang dihasilkan dari akumulasi fitur kata, semakin tinggi kemungkinan data tersebut melampaui ambang batas (threshold 0.5) untuk diklasifikasikan ke dalam kelas sentimen positif. Pada tahap implementasi komputasi, model klasifikasi ini dibangun menggunakan pustaka *Scikit-Learn* dengan pengaturan fungsi optimisasi (*solver*) ='*lbfgs*', batas iterasi maksimum (*max_iter*) =1000.

2.10 *Evaluasi Model*

Penilaian kinerja model serta untuk melihat jumlah prediksi benar dan salah diperlukan *confusion matrix* [30]. Melalui matriks ini, kinerja model dapat dievaluasi secara lebih menyeluruh. Berdasarkan Tabel 1 di bawah, nilai TP menggambarkan banyaknya data berlabel positif yang berhasil diprediksi sesuai oleh model, sedangkan TN menunjukkan jumlah data berlabel negatif yang diidentifikasi secara tepat. Nilai FP mengacu pada kesalahan ketika sampel negatif diprediksi sebagai positif, sementara FN terjadi saat sampel positif salah diklasifikasikan menjadi negatif.

Tabel 1. *Confusion Matrix*

	Kelas Aktual	
Prediksi	TP	FN
	FP	FP

Akurasi digunakan untuk menggambarkan tingkat ketepatan model dalam memprediksi kelas data secara keseluruhan, baik pada kelas positif maupun negatif, dengan membandingkan jumlah prediksi yang benar terhadap total data uji. Penghitungan akurasi ditunjukkan pada persamaan 8.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{8}$$

Presisi digunakan untuk menilai ketepatan model dalam menghasilkan prediksi positif, dengan mempertimbangkan proporsi prediksi yang benar berasal dari kelas positif. Metrik ini membantu mengurangi kemungkinan kesalahan ketika model memberikan prediksi positif yang tidak tepat. Persamaan 9 di bawah bisa digunakan dalam penghitungan presisi.

$$Precision = \frac{TP}{TP + FP} \tag{9}$$

Recall menunjukkan kemampuan model dalam menangkap seluruh data yang termasuk dalam kelas positif. Nilai recall mencerminkan seberapa besar bagian data positif yang berhasil dikenali dibandingkan dengan jumlah data positif yang tersedia. Persamaan 10 di bawah menunjukkan penghitungan *recall*.

$$Recall = \frac{TP}{TP + FN} \tag{10}$$

F1-Score digunakan sebagai ukuran gabungan antara presisi dan recall untuk memberikan penilaian yang lebih seimbang, terutama ketika kedua metrik tersebut menunjukkan nilai yang tidak sejalan. Penghitungan F1-Score ditunjukkan pada persamaan 11.

$$F1\text{-}Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{11}$$

3. HASIL DAN PEMBAHASAN

3.1 Pengumpulan Data

Proses pengumpulan data dilakukan secara otomatis menggunakan teknik *scraping* dengan memanfaatkan pustaka *YouTube Data API v3* pada bahasa pemrograman *Python*. Metode ini dipilih karena kemampuannya menyediakan akses terstruktur terhadap konten, metadata video, serta statistik interaksi pengguna secara real-time. Fokus ekstraksi data ditujukan pada kolom komentar video YouTube yang membahas mengenai kebijakan Menteri

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 4, Juni 2026
e-ISSN 2685-998X
DOI 10.30865/json.v7i4.9632

Keuangan, Bapak Purbaya Yudhi Sadewa. Melalui proses *scrapping* tersebut, berhasil dikumpulkan dataset awal (raw data) sebanyak 4.003 baris dengan 5 atribut kolom yaitu *VideoID, Author, Comment, Likes, PublishedAt*. Sampel dataset ditunjukkan pada tabel 2.

Tabel 2. Hasil Pengumpulan Data

No	VideoID	Author	Comment	Likes	PublishedAt
1.	W9Yjf-QxYmo	@Erwin-j1u	Gaspol pak menkeu	0	2025-09-
2.	W9Yjf-QxYmo	@chorulsoesety	apa sih gunanya	1	26T02:12:25Z
3.	MX1k6-rrEX	io1810	pemerintah, selain	0	2025-09-
		@ahmadmahm	tukang peras rakyat?		24T12:40:40Z
		udin1335	Semangat Mas Pur		2025-09-
					23T13:26:39Z

3.2 *Pre-processing Data*

Karakteristik data tekstual yang bersumber dari kolom komentar YouTube umumnya bersifat sangat tidak terstruktur, rentan dengan *noise*, dan sangat dipengaruhi oleh gaya bahasa warganet yang dinamis. Korpus data awal (*raw data*) yang berhasil dikumpulkan berjumlah 4.003 baris data. Data mentah ini belum dapat diproses oleh algoritma klasifikasi karena masih mengandung banyak anomali seperti variasi huruf kapital, tanda baca, angka, simbol, emoticon, serta dominasi kata-kata tidak baku. Oleh karena itu, seluruh korpus dieksekusi melalui satu *pipeline* pre-processing utuh yang secara berurutan meliputi *Case Folding, Cleaning, Normalization, Tokenizing, Stopword Removal, dan Stemming*. Transformasi ini penting agar ruang dimensi fitur (*feature space*) pada tahap TF-IDF nantinya tidak membengkak akibat pengulangan kata yang bermakna sama namun ditulis secara berbeda.

Pada fase awal, *Case Folding* menyeragamkan seluruh teks menjadi huruf kecil (*lowercase*), dilanjutkan dengan Cleaning yang mengeliminasi seluruh elemen non-alfabetik. Sebuah temuan menarik terjadi pada fase pembersihan ini; dari 4.003 baris data awal, jumlahnya menyusut menjadi 3.896 baris data bersih (*clean* data). Reduksi sebanyak 107 baris data ini disebabkan oleh adanya penghapusan data duplikat (komentar spam yang dikirim berulang) serta penghapusan baris kosong. Baris kosong ini muncul karena terdapat sejumlah pengguna yang murni hanya memberikan komentar berupa emoticon atau simbol tanpa narasi teks apa pun, sehingga ketika simbol tersebut dibersihkan, baris observasi kehilangan nilai informasinya dan harus didrop dari dataset untuk mencegah error pada komputasi model.

Setelah korpus bersih dari derau sintaksis, tantangan semantik diselesaikan melalui tahap *Normalization*. Bahasa percakapan netizen Indonesia sangat kental dengan singkatan dan bahasa gaul (*slang*). Implementasi kamus translasi kustom pada tahap ini secara signifikan mengonversi kata-kata informal menjadi padanan baku. Tanpa proses normalisasi ini, variasi penulisan seperti "tdk", "ngga", atau "engga" akan diekstrak sebagai tiga atribut kolom yang berbeda oleh TF-IDF, yang memicu *overfitting*. Teks kemudian dipecah menjadi unit kata tunggal (*Tokenizing*) sebelum disaring melalui *Stopword Removal*. Pada tahap penghapusan kata hubung ini, sebuah intervensi penting dilakukan dengan mengecualikan kata-kata negasi (seperti "tidak", "bukan", "jangan") dari daftar hapus. Penyelamatan kata negasi ini merupakan langkah preservasi konteks yang sangat esensial tanpanya, komentar kritik seperti "kinerja tidak bagus" akan kehilangan kata "tidak", sehingga berbalik makna menjadi sentimen positif saat dilabeli.

Proses pra-pemrosesan ini kemudian disempurnakan dengan algoritma Stemming menggunakan pustaka Sastrawi, yang menghapus kata imbuhan dan mereduksi setiap kata kembali ke kata dasarnya. Rangkuman komparatif yang mengilustrasikan perubahan dari data teks mentah menjadi 3.896 baris data bersih yang siap dikomputasi pada tahap pelabelan dapat dilihat pada Tabel 3.

Tabel 3. Hasil Transformasi *Pre-Processing*

Comment	After Normalize	After Tokenize	Clean Data
😎Cara orang cerdas memilih orang cerdas.Bukan karena bisikan atau suka tidak suka.Harus diuji nyali dan ketajaman intelektualitasnya,sebab keuangan memegang kunci keberhasilan atau	cara orang cerdas memilih orang cerdas bukan karena bisikan atau suka tidak suka harus diuji nyali dan ketajaman intelektualitasnya sebab keuangan memegang kunci keberhasilan atau	['cara', 'orang', 'cerdas', 'memilih', 'orang', 'cerdas', 'bukan', 'karena', 'bisikan', 'atau', 'suka', 'tidak', 'suka', 'harus', 'diuji', 'nyali', 'dan', 'ketajaman', 'intelektualitasnya', 'sebab', 'keuangan', 'memegang', 'kunci',	orang cerdas pilih orang cerdas bisik suka suka uji nyali tajam intelektualitasnya uang pegang kunci hasil gagal perintah

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 4, Juni 2026
e-ISSN 2685-998X
DOI 10.30865/json.v7i4.9632

gagalnya suatu pemerintahan. 👍👍👍	gagalnya suatu pemerintahan	'keberhasilan', 'atau', 'gagalnya', 'suatu', 'pemerintahan']	
Lajut bpk purba. Elitnya yg g bayr pajk. Paksa bayar pajk kita rakyat Indonesia mendukung	lanjut bapak purba elitnya yang tidak bayar pajak paksa bayar pajak kita rakyat indonesia mendukung	['lanjut', 'bapak', 'purba', 'elitnya', 'yang', 'tidak', 'bayar', 'pajak', 'paksa', 'bayar', 'pajak', 'kita', 'rakyat', 'indonesia', 'mendukung']	purba elitnya bayar pajak paksa bayar pajak rakyat indonesia dukung
Iya jangan rakyat kecil ajh yg di kejar.dan di ancem2 sama pajak	iya jangan rakyat kecil aja yang di kejar dan di ancam sama pajak	['iya', 'jangan', 'rakyat', 'kecil', 'aja', 'yang', 'di', 'kejar', 'dan', 'di', 'ancam', 'sama', 'pajak']	iya rakyat aja kejar ancam pajak

3.3 Pelabelan Data

Tahapan pelabelan sentimen dieksekusi terhadap 3.896 data bersih hasil *pre-processing*. Mengacu pada aturan keputusan (*decision rule*) berbasis *Indonesian Sentiment Lexicon* (InSet) yang telah ditetapkan pada fase metodologi, setiap komentar dikuantifikasi bobot polaritasnya dan dipetakan ke dalam kelas sentimen (positif & negatif) secara otomatis.

Setelah proses pelabelan komputasional ini diterapkan secara komprehensif pada seluruh korpus, hasil kalkulasi mengungkap adanya disparitas proporsi yang sangat ekstrem antara komentar bernada positif dan negatif. Sebaran ini memberikan gambaran kuantitatif awal mengenai kecenderungan opini publik terhadap manuver dan kebijakan Menteri Keuangan pada platform YouTube. Berdasarkan hasil klasifikasi, narasi publik didominasi secara masif oleh sentimen negatif yang mencapai 3.002 baris data. Di sisi lain, sentimen positif hanya merepresentasikan kelompok minoritas sejumlah 894 baris data. Guna memberikan gambaran mengenai bagaimana sistem mengevaluasi teks berdasarkan leksikon tersebut, representasi beberapa sampel data komentar beserta hasil pelabelan sentimen akhirnya ditunjukkan pada Tabel 4.

Tabel 4. Contoh Hasil Pelabelan Data

No	clean data	label
1	sih guna perintah tukang peras rakyat	negatif
2	prabowo purbaya indonesia jaya gas	positif
3	mantab purbaya menteri peka situasi kondisi alami rakyat	positif
4	tahu agustus demo protes rakyat susah cari kerja warung sepi harga sembako mahal	negatif

3.4 Hasil Ekstraksi Fitur (TF-IDF)

Setelah tahap pembersihan data selesai, 3.896 data komentar teks yang tersisa perlu diubah menjadi format angka agar dapat dipahami dan diproses oleh algoritma komputer. Proses pengubahan ini menggunakan metode pembobotan *Term Frequency-Inverse Document Frequency* (TF-IDF). Hasilnya, sistem berhasil mendeteksi dan mengekstrak 5.000 kata unik dari keseluruhan komentar. Metode TF-IDF ini memastikan bahwa kata yang sekadar sering muncul seperti kata hubung yang mungkin terlewat tidak akan mendominasi, melainkan memberikan bobot tinggi pada kata-kata yang benar-benar unik dan memiliki makna penting. Untuk melihat gambaran topik apa saja yang paling menonjol dalam opini publik terhadap Menteri Keuangan, 10 kata dengan bobot TF-IDF tertinggi disajikan pada Tabel 5.

Tabel 5. Hasil Ekstraksi Fitur 10 Teratas

No	kata	TF-IDF
1	menteri	0.029778
2	rakyat	0.024864
3	moga	0.021871
4	uang	0.019952
5	ganti	0.017889
6	orang	0.017721
7	purbaya	0.017352
8	ekonomi	0.016165
9	menkeu	0.016119

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 4, Juni 2026
e-ISSN 2685-998X
DOI 10.30865/json.v7i4.9632

No	kata	TF-IDF
10	indonesia	0.014990

Berdasarkan Tabel 5, terlihat jelas bahwa kata-kata yang mendapat skor tertinggi sangat berkaitan erat dengan isu ekonomi dan kebijakan, seperti kata "menteri" dan rakyat". Munculnya kata-kata spesifik ini membuktikan bahwa metode TF-IDF berhasil menangkap inti pembicaraan dan keresahan publik dengan sangat baik. Hasil pembobotan kata ini memiliki peran yang sangat krusial untuk tahapan selanjutnya. Deretan angka dari TF-IDF ini akan menjadi petunjuk utama bagi algoritma *Support Vector Machine* dan *Logistic Regression* dalam mengklasifikasikan sentimen.

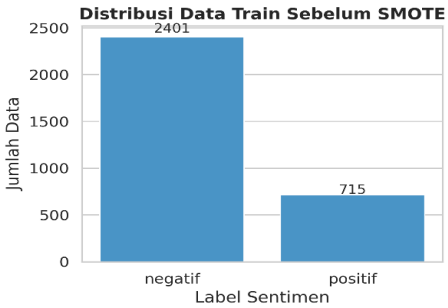
3.5 *Handling Imbalance Data*

Sebelum menangani masalah ketimpangan kelas, dataset bersih yang berjumlah 3.896 baris terlebih dahulu dipisah menjadi dua bagian dengan rasio 80:20, yakni data latih sebanyak 3.116 baris dan data uji sebanyak 780 baris. Tahapan penyeimbangan data ini wajib diimplementasikan secara eksklusif hanya pada data latih agar objektivitas evaluasi pada data uji nantinya tetap terjaga.

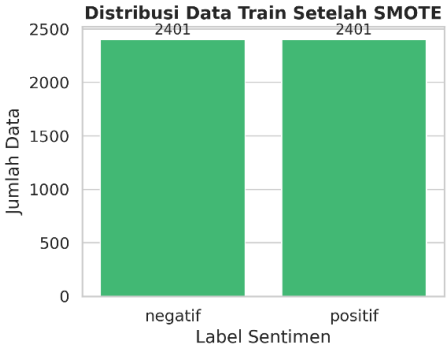
Jika melihat distribusi pada 3.116 data latih tersebut, terlihat jelas adanya ketidakseimbangan jumlah sampel yang sangat tajam. Sentimen negatif mendominasi secara masif dengan total 2.401 komentar, sedangkan sentimen positif yang mewakili dukungan publik hanya berjumlah 715 komentar. Jika kondisi timpang ini dibiarkan begitu saja, algoritma *machine learning* nantinya akan menjadi bias. Model akan cenderung selalu menebak sentimen negatif karena jumlahnya jauh lebih banyak. Akibatnya, model akan kehilangan kepekaan untuk mengenali pola-pola pada kelas minoritas, sehingga aspirasi publik yang bernada positif terhadap kebijakan Menteri Keuangan justru berisiko terabaikan oleh sistem.

Untuk mengatasi ancaman bias tersebut, teknik *Synthetic Minority Over-sampling Technique* (SMOTE) diterapkan. Alih-alih sekadar menggandakan data positif yang sudah ada yang bisa memicu *overfitting*, SMOTE bekerja lebih cerdas dengan cara mempelajari pola karakteristik dari 715 komentar positif tersebut, lalu menciptakan sampel data buatan baru yang mirip. Melalui intervensi ini, jumlah kelas positif yang awalnya tertinggal jauh berhasil didongkrak hingga menyamai kelas mayoritas, yakni menjadi 2.401 data.

Setelah proses SMOTE selesai dieksekusi, komposisi data latih kini berubah drastis menjadi seimbang sempurna dengan proporsi 50:50. Ruang data yang sudah proporsional ini memberikan kondisi belajar yang sangat ideal dan adil bagi algoritma pemodelan. Model kini siap dilatih untuk mengenali karakteristik kedua sentimen tanpa ada satu kelas pun yang didominasi oleh kelas lainnya. Visualisasi yang memperlihatkan perbandingan distribusi kelas sebelum dan sesudah penerapan SMOTE ini ditampilkan secara jelas pada Gambar 2 dan Gambar 3.



Gambar 2. Data Train Sebelum SMOTE



Gambar 3. Data Train Setelah SMOTE

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 4, Juni 2026
e-ISSN 2685-998X
DOI 10.30865/json.v7i4.9632

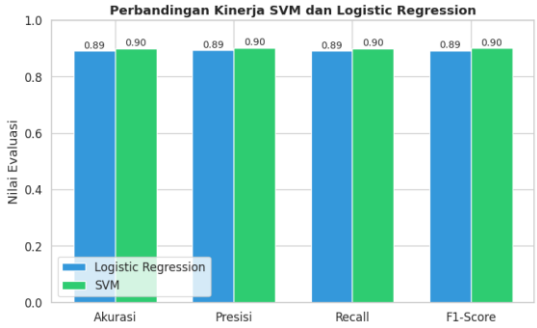
3.6 Pengujian dan Evaluasi Model

Berdasarkan pengujian komparatif pada 780 data uji, algoritma *Support Vector Machine* terbukti sedikit mengungguli *Logistic Regression*. Perbandingan hasil evaluasi secara menyeluruh divisualisasikan pada Tabel 11 dan Gambar 4.

Tabel 11. Perbandingan Performa Model

No	Metriks	SVM	Logistic Regression
1.	Akurasi	90 %	89,10 %
2.	Presisi	90,18 %	89,32 %
3.	Recall	90 %	89,10 %
4.	F1-Score	90,08 %	89,20 %

Untuk memperjelas perbandingan kinerja kedua model tersebut, hasil evaluasi disajikan dalam bentuk grafik pada gambar 4.



Gambar 4. Perbandingan Kinerja SVM dan Logistic Regression

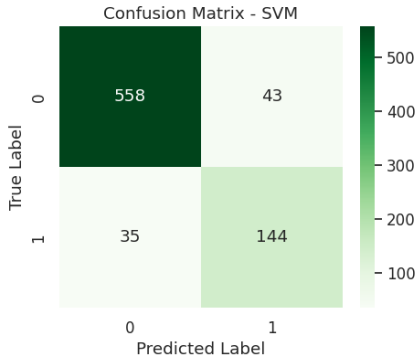
Merujuk pada Tabel 11, SVM memimpin dengan capaian akurasi sebesar 90%, berbanding 89,10% pada LR. Keunggulan ini secara teoretis dapat dijelaskan melalui karakteristik data teks itu sendiri. Ekstraksi fitur menggunakan TF-IDF menghasilkan matriks berdimensi tinggi. SVM, khususnya dengan pendekatan margin maksimal pada pemisah liniernya, sangat tangguh (robust) dalam menangani data berdimensi tinggi tersebut tanpa mudah terjebak pada overfitting. Di sisi lain, Logistic Regression yang berbasis pada estimasi probabilitas cenderung sedikit lebih sensitif terhadap keberadaan outlier pada ruang fitur.

Lebih lanjut, dampak krusial dari penerapan SMOTE terlihat jelas pada capaian metrik kelas minoritas. Sebelum SMOTE diterapkan, model diasumsikan akan memiliki tingkat Recall yang sangat rendah pada kelas minoritas (positif) akibat bias kelas mayoritas (negatif). Dengan intervensi SMOTE, wilayah keputusan (decision boundary) dipaksa melebar secara adil. Hal ini terbukti dari nilai F1-Score kelas positif yang mencapai 79%, mengindikasikan bahwa model tidak lagi mengabaikan narasi dukungan publik terhadap kebijakan Menteri Keuangan, melainkan mampu mendeteksinya dengan tingkat presisi dan sensitivitas yang proporsional.

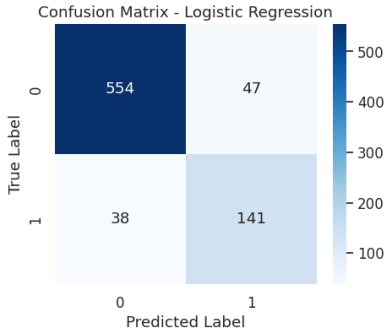
Selain penyajian hasil dalam bentuk tabel dan grafik perbandingan, penelitian ini juga memanfaatkan confusion matrix untuk membedah lebih dalam bagaimana model melakukan prediksi pada masing-masing kelas

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 4, Juni 2026
e-ISSN 2685-998X
DOI 10.30865/json.v7i4.9632

secara spesifik di lapangan, baik yang sesuai maupun yang keliru. Representasi confusion matrix untuk kedua modelditampilkan pada Gambar 5 dan Gambar 6.



Gambar 5. Confusion Matrix SVM



Gambar 6. Confusion Matrix Logistic Regression

Berdasarkan representasi pada Gambar 5, model SVM secara spesifik berhasil menebak 702 observasi data dengan presisi tinggi, yakni kombinasi 558 komentar negatif (*True Negative*) dan 144 komentar positif (*True Positive*). Kemampuan model mengenali 144 dari total 179 data minoritas ini menjadi bukti konkret keberhasilan teknik SMOTE di tahap sebelumnya. Namun demikian, matriks tersebut juga mengungkap sisi kelemahan model saat melakukan klasifikasi. Terdapat 43 komentar negatif yang salah ditebak sebagai positif (*False Positive*), dan 35 komentar positif yang luput lalu ditebak sebagai negatif (*False Negative*).

Di sisi lain, representasi pada Gambar 6 menunjukkan bahwa model *Logistic Regression* (LR) menghasilkan kinerja yang sedikit di bawah SVM. Model LR berhasil menebak 695 observasi data dengan benar, yang mencakup 554 komentar negatif (*True Negative*) dan 141 komentar positif (*True Positive*). Meskipun mampu mengenali 141 data minoritas, model LR memiliki tingkat kesalahan prediksi yang sedikit lebih tinggi dibandingkan SVM, yakni terdapat 47 komentar negatif yang salah ditebak sebagai positif (*False Positive*) serta 38 komentar positif yang gagal dideteksi dan ditebak sebagai negatif (*False Negative*).

Kegagalan prediksi pada puluhan data tersebut bersumber dari kelemahan arsitektur eksperimen ini, yakni ketergantungan pada pendekatan *Lexicon-based* (InSet) sebagai penentu ground truth yang rentan terhadap anomali semantik. Model yang dilatih berdasarkan pelabelan ini mewarisi keterbatasan leksikon dalam mendeteksi sarkasme, ironi, atau nuansa bahasa Indonesia yang kompleks. Sebuah komentar cemoohan berbalut pujian palsu akan tetap dianggap positif oleh model. Selain itu, metode TF-IDF juga membuat kalimat kehilangan konteks urutan kata, sehingga letak kata negasi yang terpisah jauh dari kata utamanya dapat mengecoh pemahaman sistem.

Terlepas dari sejumlah keterbatasan tersebut, penelitian ini memvalidasi bahwa kombinasi SMOTE dengan algoritma klasifikasi SVM maupun *Logistic Regression* merupakan *pipeline* yang efektif untuk memitigasi ketimpangan kelas pada data opini publik. Secara praktis, model komputasi yang dikembangkan dapat diimplementasikan sebagai sistem *early warning* atau instrumen analitik otomatis bagi pemangku kebijakan. Sistem ini memungkinkan ekstraksi sentimen publik yang masif dan langsung (*real-time*) tanpa perlu melakukan pelabelan manual yang memakan waktu, sehingga respons masyarakat terhadap kebijakan atau figur Menteri Keuangan dapat dipetakan secara objektif dan terukur.

4. KESIMPULAN

Berdasarkan penelitian yang telah dilakukan untuk memetakan persepsi publik terhadap Menteri Keuangan Republik Indonesia melalui platform YouTube, dapat disimpulkan bahwa implementasi algoritma machine

learning berhasil melakukan klasifikasi dengan sangat baik. Secara spesifik, *Support Vector Machine* menunjukkan performa yang sedikit lebih unggul yaitu dengan tingkat akurasi 90%, dibandingkan dengan *Logistic Regression* yang menunjukkan akurasi sebesar 89% dalam klasifikasi sentimen komentar YouTube. Capaian akurasi yang tinggi pada kedua model ini tidak terlepas dari peran krusial penerapan teknik *Synthetic Minority Over-sampling Technique (SMOTE)*. Penggunaan SMOTE terbukti efektif dalam menangani masalah ketidakseimbangan kelas (*class imbalance*) pada dataset komentar, sehingga model terhindar dari bias prediksi terhadap kelas mayoritas. Jika dikaitkan dengan temuan penelitian terdahulu, hasil ini secara empiris mengindikasikan adanya peningkatan kinerja model yang mencerminkan efektivitas tahapan pengolahan data, pembobotan TF-IDF, dan strategi pemodelan yang diterapkan. Kontribusi utama penelitian ini terletak pada pemanfaatan komentar YouTube yang panjang dan argumentatif sebagai sumber analisis sentimen terhadap figur pejabat publik, serta pada capaian performa model yang lebih baik dibandingkan penelitian sebelumnya dengan konteks dan *platform* data yang berbeda. Meskipun demikian, penelitian ini masih memiliki keterbatasan, terutama pada penggunaan ekstraksi fitur berbasis frekuensi kata yang terkadang belum sepenuhnya mampu menangkap makna semantik atau ironi dari teks yang tidak baku. Oleh karena itu, untuk pengembangan penelitian selanjutnya, penggunaan teknik ekstraksi fitur yang lebih kompleks, seperti word embeddings misalnya *Word2Vec* atau *GloVe*, serta penggunaan dataset yang lebih besar dapat dipertimbangkan guna melihat potensi peningkatan performa pada data komentar YouTube yang memiliki karakteristik bahasa yang sangat beragam.

REFERENCES

[1] Kementerian Keuangan Republik Indonesia, "Serah Terima Jabatan Menteri Keuangan Republik Indonesia." Accessed: Oct. 11, 2025. [Online]. Available: https://www.kemenkeu.go.id/informasi-publik/publikasi/berita-utama/Sertijab-Menkeu-RI

[2] L. Rahayu, "Profil Purbaya Yudhi Sadewa, Menkeu Baru Pengganti Sri Mulyani," Detik.com. Accessed: Oct. 11, 2025. [Online]. Available: https://news.detik.com/berita/d-8101667/profil-purbaya-yudhi-sadewa-menkeu-baru-pengganti-sri-mulyani

[3] Tempo.co, "Kenapa Pasar Langsung Bereaksi Iringi Pergantian Sri Mulyani ke Purbaya | tempo.co." Accessed: Oct. 12, 2025. [Online]. Available: https://www.tempo.co/ekonomi/kenapa-pasar-langsung-bereaksi-iringi-pergantian-sri-mulyani-ke-purbaya-2068013

[4] R. F. Alhujaili and W. M. S. Yafooz, "Sentiment Analysis for YouTube Educational Videos Using Machine and Deep Learning Approaches," in *2022 IEEE 2nd International Conference on Electronic Technology, Communication and Information (ICETCI)*, 2022, pp. 238–244. doi: 10.1109/ICETCI55101.2022.9832284.

[5] S. Kemp, "Pengguna TikTok, Statistik, Data, Tren, dan Lainnya — DataReportal – Wawasan Digital Global," Data Reportal. Accessed: Oct. 16, 2025. [Online]. Available: https://datareportal.com/essential-youtube-stats

[6] P. Chauhan, N. Sharma, and G. Sikka, "The emergence of social media data and sentiment analysis in election prediction," *J. Ambient Intell. Humaniz. Comput.*, vol. 12, no. 2, pp. 2601–2627, 2021, doi: 10.1007/s12652-020-02423-y.

[7] V. Joseph, C. P. Lora, and N. T, "Exploring the Application of Natural Language Processing for Social Media Sentiment Analysis," *2024 3rd Int. Conf. Innov. Technol.*, pp. 1–6, 2024, [Online]. Available: https://api.semanticscholar.org/CorpusID:269622880

[8] M. Bordoloi and S. K. Biswas, "Sentiment analysis: A survey on design framework, applications and future scopes," *Artif. Intell. Rev.*, vol. 56, no. 11, pp. 12505–12560, 2023, doi: 10.1007/s10462-023-10442-2.

[9] M. Wankhade, A. C. S. Rao, and C. Kulkarni, "A survey on sentiment analysis methods, applications, and challenges," *Artif. Intell. Rev.*, vol. 55, no. 7, pp. 5731–5780, Oct. 2022, doi: 10.1007/s10462-022-10144-1.

[10] R. Rodríguez-Pérez and J. Bajorath, "Evolution of Support Vector Machine and Regression Modeling in Chemoinformatics and Drug Discovery," *J. Comput. Aided. Mol. Des.*, vol. 36, no. 5, pp. 355–362, 2022, doi: 10.1007/s10822-022-00442-9.

[11] P. Rajendra and S. Latifi, "Prediction of diabetes using logistic regression and ensemble techniques," *Comput. Methods Programs Biomed. Updat.*, vol. 1, p. 100032, 2021, doi: https://doi.org/10.1016/j.cmpbup.2021.100032.

[12] N. Hadi and D. Sugiarto, "Analisis Sentimen Pembangunan IKN pada Media Sosial X Menggunakan Algoritma SVM, Logistic Regression dan Naïve Bayes," *J. Inform. J. Pengemb. IT*, vol. 10, no. 1, pp. 37–49, 2025, doi: 10.30591/jpit.v10i1.7106.

[13] A. Maulana, Inayah Khasnaputri Afifah, Asghafi Mubarrak, Kiagus Rachmat Fauzan, Ardhan Dwintara, and B. P. Zen, "Comparison of Logistic Regression, Multinomialnb, Svm, and K-Nn Methods on Sentiment Analysis of Gojek App Reviews on the Google Play Store," *J. Tek. Inform.*, vol. 4, no. 6, pp. 1487–1494, Dec. 2023, doi: 10.52436/1.jutif.2023.4.6.863.

[14] I. R. Ainunnisa, "Analisis Sentimen Aplikasi Tiktok dengan Metode Support Vector Machine (SVM), Logistic Regression dan Naïve Bayes," vol. 6, no. 3, pp. 423–430, 2023, doi: 10.32493/jtsi.v6i3.31076.

[15] Z. C. Dwinnie and R. Novita, "Penerapan Machine Learning Pada Analisis Sentimen Twitter Sebelum dan Sesudah Debat Calon Presiden dan Wakil Presiden Tahun 2024," vol. 8, no. April, pp. 758–767, 2024, doi: 10.30865/mib.v8i2.7504.

[16] R. A. Prasetyo, "Perbandingan Algoritma Logistic Regression, SVM, dan Random Forest pada Analisis Sentimen Aplikasi Gopay," *J. Inform. J. Pengemb. IT*, vol. 10, no. 4, pp. 1176–1188, 2025, doi: 10.30591/jpit.v10i4.8796.

[17] D. Elreedy, A. F. Atiya, and F. Kamalov, "A theoretical distribution analysis of synthetic minority oversampling technique (SMOTE) for imbalanced learning," *Mach. Learn.*, pp. 1–21, 2023, [Online]. Available: https://api.semanticscholar.org/CorpusID:255728052

[18] N. Chawla, K. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *ArXiv*, vol. abs/1106.1, 2002, [Online]. Available: https://api.semanticscholar.org/CorpusID:1554582

[19] M. A. Jassim and S. N. Abdulwahid, "Data Mining preparation: Process, Techniques and Major Issues in Data Analysis," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 1090, no. 1, p. 12053, Mar. 2021, doi: 10.1088/1757-899X/1090/1/012053.

[20] H. T. Duong and T. A. Nguyen-Thi, "A review: preprocessing techniques and data augmentation for sentiment analysis," *Comput. Soc. Networks*, vol. 8, no. 1, pp. 1–16, 2021, doi: 10.1186/s40649-020-00080-x.

[21] K. S. Prathyusha and B. E. Reddy, "Normalization methods for multiple sources of data," *Proc. - 5th Int. Conf. Intell. Comput. Control Syst. ICICCS 2021*, no. Iciccs, pp. 1013–1019, 2021, doi: 10.1109/ICICCS51141.2021.9432142.

[22] P. Vadlapati, "Tokenization Beyond NLP: Potential Applications in Data Analytics, Cybersecurity, and Beyond," *Interantional J. Sci. Res. Eng. Manag.*, vol. 08, no. 12, pp. 1–7, 2024, doi: 10.55041/ijsrem9532.

[23] D. A. N. Arifin, S. Pada, D. Teks, B. Indonesia, J. Pardede, and D. Darmawan, "Perbandingan Algoritma Stemming Porter , Sastrawi , Idris , Comparison of Stemming Algorithms Porter , Sastrawi , Idris , and Arifin Setiono on Indonesian Text Documents," vol. 12, no. 1, 2025, doi: 10.25126/jtiik.2025128860.

[24] Fajri Koto and Gemala Y. Rahmaningtyas, "InSet Lexicon: Evaluation of a Word List for Indonesian Sentiment Analysis in Microblogs," *2017 Int. Conf. Asian Lang. Process.*, pp. 391–394, 20AD.

[25] A. Ma, A. E. Pajri, and P. Liana, "Sentiment Analysis of President Prabowo ' s Performance on Twitter (X) with a Comparative Study of SVM , XGBoost , and AdaBoost," vol. 10, no. 1, pp. 684–697, 2026.

[26] I. S. Wibowo, A. Witanti, and I. Susilawati, "Keyword Extraction Judul Berita Online Di Indonesia Menggunakan Metode TF-IDF," *J. Tek. Inform. dan Sist. Inf.*, vol. 11, no. 1, pp. 99–111, 2024, [Online]. Available: http://jurnal.mdp.ac.id

[27] W. A. N. G. Zhuohao, W. A. N. G. Dong, and L. I. Qing, "Keyword Extraction from Scientific Research Projects Based on SRP-TF-IDF," *Chinese J. Electron.*, vol. 30, no. 4, pp. 652–657, 2021, doi: 10.1049/cje.2021.05.007.

[28] 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.

[29] I. Akbar and M. Faisal, "Perbandingan Analisis Sentimen PLN Mobile: Machine Learningvs. Deep Learning," vol. 7, no. 1, 2022, doi: https://doi.org/10.31328/jointecs.v8i1.5078.

[30] J. H. Cabot and E. G. Ross, "Evaluating prediction model performance," *Surgery*, vol. 174, no. 3, pp. 723–726, Sep. 2023, doi: 10.1016/j.surg.2023.05.023.