Building of Informatics, Technology and Science (BITS)
Volume 7, No 3, December 2025 Page: 1512−1524
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i3.8513

Analisis Sentimen Komentar YouTube terhadap Kenaikan Tunjangan DPR RI menggunakan Naïve Bayes, SVM, dan Random Forest

Jemmi Rama Dani[1,*], Parjito[2]

[1] Fakultas Teknik dan Ilmu Komputer, Program Studi Sistem Informasi, Universitas Teknokrat Indonesia, Bandar Lampung, Indonesia
[2] Fakultas Teknik dan Ilmu Komputer, Program Studi Teknik Komputer, Universitas Teknokrat Indonesia, Bandar Lampung, Indonesia
Email: [1,*]jemmi_rama_dani@teknokrat.ac.id, [2]djito@teknokrat.ac.id
Email Penulis Korespondensi: jemmi_rama_dani@teknokrat.ac.id
Submitted: **11/10/2025**; Accepted: **06/12/2025**; Published**: 08/12/2025**

Abstrak−Perkembangan teknologi digital mendorong masyarakat untuk lebih aktif menyuarakan opini melalui media sosial, termasuk dalam menanggapi isu politik seperti kebijakan kenaikan tunjangan DPR RI. Penelitian ini bertujuan untuk menganalisis sentimen publik terhadap isu tersebut pada platform *YouTube* dengan pendekatan komparatif menggunakan tiga algoritma *Machine Learning*, yaitu *Naïve Bayes, Support Vector Machine*, dan *Random Forest*. Data diperoleh dari komentar penonton melalui *YouTube* Data *Application Programming Interface* (API) dengan total sebanyak 78.866 baris komentar yang dikumpulkan dari tujuh video yang membahas kontroversi DPR RI. Proses pengambilan data dilakukan menggunakan modul *googleapiclient.discovery.build* dengan API versi V3, di mana API_Key berfungsi sebagai kunci *autentikasi* untuk mengakses data dari *YouTube.* Tahapan penelitian mencakup *preprocessing* untuk pembersihan data, pelabelan sentimen berbasis *InSet lexicon Based*, serta penerapan *Synthetic Minority Over-sampling Technique* (SMOTE) guna mengatasi ketidakseimbangan data antar kelas. Hasil penelitian menunjukkan bahwa sebelum penerapan SMOTE, model *Support Vector Machine* (SVM) mencapai akurasi tertinggi sebesar 89%, diikuti oleh *Random Forest* dengan 81%, dan *Naïve Bayes* dengan 62%. Setelah penerapan SMOTE, kinerja ketiga model meningkat signifikan, dengan SVM memperoleh akurasi tertinggi 93%, disusul *Random Forest* sebesar 86%, dan *Naïve Bayes* sebesar 75%. Pada kelas positif, SVM juga menunjukkan performa terbaik dengan nilai *Precision* 96% dan *Recall* 95%, dengan *F1-Score* sebesar 95%. Secara keseluruhan, hasil penelitian ini menegaskan bahwa SVM lebih unggul dalam menjaga keseimbangan klasifikasi antar kelas, baik sebelum maupun sesudah SMOTE. Pendekatan analisis sentimen berbasis *Machine Learning* terbukti mampu memberikan gambaran yang komprehensif mengenai opini publik terhadap isu politik, sekaligus menjadi masukan penting bagi pembuat kebijakan dalam merumuskan strategi komunikasi yang lebih transparan dan responsif.

Kata Kunci: Analisis Sentimen; DPR RI; YouTube; Support Vector Machine; Naïve Bayes; SMOTE

Abstract−The rise of digital technology encourages the public to actively voice their opinions through social media, including in response to political issues such as the policy on increasing the remuneration of the Indonesian House of Representatives (DPR RI). This research aims to analyze public sentiment towards this issue on the YouTube platform using a comparative approach with three Machine Learning algorithms: Naïve Bayes, Support Vector Machine, and Random Forest. The data was acquired from viewer comments via the YouTube Data Application Programming Interface (API), totaling 78,866 lines of comments collected from seven videos discussing the DPR RI controversy. The data collection process utilized the googleapiclient.discovery.build module with API version V3, where the API_Key served as the authentication key to access data from YouTube. The research stages included preprocessing for data cleaning, sentiment labeling based on the InSet Lexicon Based method, and the application of the Synthetic Minority Over-sampling Technique (SMOTE) to address class imbalance in the data. The results show that before SMOTE application, the Support Vector Machine (SVM) model achieved the highest accuracy of 89%, followed by Random Forest at 81%, and Naïve Bayes at 62%. After applying SMOTE, the performance of all three models increased significantly, with SVM obtaining the highest accuracy of 93%, followed by Random Forest at 86%, and Naïve Bayes at 75%. For the positive class, SVM also demonstrated the best performance with a Precision value of 96%, Recall of 95%, and an F1-Score of 95%. Overall, the findings of this study confirm that SVM is superior in maintaining class balance in classification, both before and after SMOTE. The Machine Learning-based sentiment analysis approach is proven capable of providing a comprehensive overview of public opinion on political issues, while also offering important input for policymakers in formulating more transparent and responsive communication strategies.

Keywords: Sentiment Analysis; DPR RI; YouTube; Support Vector Machine; Naïve Bayes; SMOTE

1. PENDAHULUAN

Indonesia merupakan negara demokrasi, hal ini ditunjukkan dengan dilakukannya pemilihan umum untuk memilih kepala negara, kepala daerah serta badan legislatif(DPR) [1]. DPR merupakan suatu lembaga negara yang bergerak didalam lingkup politik hukum, dan Undang-Undang sebagai manifestasi dari politik tersebut. DPR memiliki kekuasaan sebagai pembentuk UndangUndang yang telah diatur dalam Undang-Undang Dasar Republik Indonesia Tahun 1945 pasal 20 ayat 1 [2]. DPR sebagai lembaga yang mewakilkan rakyat menyandang tanggung jawab yang harusnya dipenuhi secara demokratis dan responsif untuk mengatasi permasalahan sosial yang ada pada masyarakat serta tidak memprioritaskan kepentingan partai politik [3]. Isu kenaikan tunjangan DPR RI telah memicu perhatian luas di tengah masyarakat karena dinilai berhubungan erat dengan prinsip keadilan sosial, efektivitas serta transparansi pengelolaan anggaran negara, dan tingkat kepercayaan publik terhadap integritas serta legitimasi lembaga legislatif sebagai representasi rakyat [4]. Salah satu platform yang banyak digunakan masyarakat untuk menyampaikan opini terkait isu ini adalah *YouTube*, di mana komentar warganet dapat mencerminkan pandangan publik secara lebih spontan dan beragam. Komentar tersebut tidak hanya digunakan sebagai sarana ekspresi, tetapi juga dapat menjadi

Building of Informatics, Technology and Science (BITS)
Volume 7, No 3, December 2025 Page: 1512−1524
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i3.8513

sumber data untuk dianalisis lebih lanjut guna mengetahui kecenderungan sentimen masyarakat [5]. Namun, persepsi publik terhadap kebijakan ini masih beragam. Sebagian pihak mendukung alasan peningkatan kinerja dan kesejahteraan anggota DPR, sementara sebagian lainnya mengkritik karena dianggap membebani anggaran dan tidak berpihak pada rakyat. Permasalahan yang muncul adalah kurangnya pemahaman mengenai opini publik terbentuk dan seberapa besar dominasi sentimen positif, negatif, maupun netral dalam menanggapi isu ini. Tantangan ini semakin kompleks karena terjadi di era digital yang sarat dengan opini, informasi, serta perdebatan yang sangat dinamis.

YouTube menjadi salah satu ruang utama bagi masyarakat untuk menyampaikan pendapat, berbagi pengalaman, serta mendiskusikan isu-isu terkait kebijakan publik, termasuk mengenai kenaikan tunjangan DPR RI. Hingga tahun 2023 , jumlah pengguna aktif *YouTube* di Indonesia mencapai lebih dari 139 juta, menempatkannya sebagai platform media sosial dengan basis pengguna terbesar di tanah air dan sekaligus menjadikannya sebagai sumber potensial untuk memantau opini publik [6]. Komentar yang muncul pada video terkait isu kenaikan tunjangan DPR RI mencerminkan beragam sentimen, mulai dari dukungan hingga kritik keras terhadap kebijakan tersebut . Dengan sifatnya yang terbuka, interaktif, dan kaya akan data teks, *YouTube* menghadirkan sumber data yang berharga untuk memahami pola diskusi, tren opini, serta persepsi masyarakat terhadap kebijakan legislatif. Akan tetapi, banyaknya komentar yang bersifat spontan dan subjektif juga meningkatkan risiko bias serta kesulitan dalam mengidentifikasi kecenderungan sentimen secara objektif [7].

Solusi yang diusulkan dalam penelitian ini adalah melakukan analisis sentimen komentar *YouTube* untuk mengidentifikasi pola opini masyarakat terhadap isu kenaikan tunjangan DPR RI . Analisis ini memanfaatkan tiga metode klasifikasi yang umum digunakan dalam pengolahan teks, yaitu *Naïve Bayes, Support Vector Machine* (SVM), dan *Random Forest*. Metode SVM bekerja dengan mencari *hyperplane* optimal yang dapat memisahkan data berdasarkan kelas sentimen, sehingga sangat sesuai untuk menangani data komentar yang berdimensi tinggi[8]. Sementara itu, *Naïve Bayes* memanfaatkan prinsip *probabilitas Bayes* dengan asumsi independensi antar fitur, sehingga mampu melakukan klasifikasi teks dengan cepat dan efisien[9]. Adapun *Random Forest* menggabungkan banyak pohon keputusan dalam proses voting untuk menghasilkan prediksi yang lebih stabil dan akurat [10]. Dengan menerapkan ketiga metode ini, penelitian diharapkan dapat memberikan pemahaman yang lebih mendalam mengenai persepsi publik terhadap kebijakan kenaikan tunjangan DPR RI, sekaligus menjadi dasar bagi pengambilan keputusan serta strategi komunikasi yang lebih tepat.

Penelitian sebelumnya telah membuktikan relevansi penerapan analisis sentimen dengan metode *Naïve Bayes*, SVM, dan *Random Forest*. Penelitian Ardiansyah et al. menemukan bahwa kombinasi *Naïve Bayes* dan *Random Forest* mampu memberikan hasil yang cukup baik dalam menganalisis komentar *YouTube* terkait eSIM, dengan *Random Forest* menunjukkan performa lebih stabil pada data berdimensi tinggi dan akurasi mencapai 87%. Selanjutnya, Syafia et al. membandingkan SVM dan *Random Forest* pada analisis komentar *YouTube* dan menemukan bahwa SVM unggul dalam hal akurasi, yaitu sebesar 85%, ketika menangani dataset besar dengan variasi opini yang kompleks. Khomsah menyoroti penggunaan *Word2Vec* dan *Random Forest* untuk memahami sentimen publik di *YouTube*, yang menegaskan efektivitas metode *ensemble* dalam menangani teks pendek dan bervariasi, dengan akurasi sebesar 82%. Triana Putri et al. juga menerapkan *Naïve Bayes* pada komentar pengguna *YouTube,* dengan hasil akurasi 80% yang menunjukkan efisiensi metode ini untuk dataset kecil hingga menengah dengan waktu komputasi yang cepat. Terakhir, Amly et al. di Indonesia menggunakan *Naïve Bayes* untuk menganalisis komentar terkait kebijakan pendidikan, dan hasilnya menunjukkan tingkat akurasi sebesar 83% serta relevan dalam konteks lokal.

Berdasarkan penelitian-penelitian tersebut, terlihat bahwa baik SVM, *Naïve Bayes*, maupun *Random Forest* telah terbukti efektif untuk analisis sentimen di berbagai bidang, mulai dari isu kesehatan, ulasan produk, kebijakan publik, hingga opini masyarakat di media sosial. Namun, gap penelitian masih terlihat jelas dalam konteks penerapannya pada analisis sentimen publik terkait isu politik di Indonesia, khususnya mengenai respons masyarakat terhadap kebijakan kenaikan tunjangan DPR RI [11]. Hingga saat ini, studi yang secara khusus membahas perbandingan performa ketiga metode tersebut dalam menganalisis opini publik pada platform *YouTube* masih sangat terbatas. Padahal, memahami pola sentimen masyarakat di ranah digital sangat penting untuk mendukung transparansi, evaluasi kebijakan, serta perumusan strategi komunikasi publik yang lebih efektif. Oleh karena itu, penelitian ini bertujuan untuk menganalisis sentimen publik terhadap kenaikan tunjangan DPR RI melalui komentar pada platform *YouTube* dengan menggunakan metode *Naïve Bayes, Support Vector Machine* (SVM), dan *Random Forest*. Penelitian ini juga akan membandingkan kinerja ketiga metode tersebut dalam aspek akurasi, presisi, *recall* dan *F1-Score.* Hasil penelitian diharapkan dapat memberikan gambaran yang lebih jelas mengenai persepsi publik terhadap kebijakan kenaikan tunjangan DPR RI, menjadi dasar dalam perumusan strategi komunikasi publik yang lebih efektif, serta memberikan kontribusi bagi literatur akademik mengenai penerapan analisis sentimen pada isu politik di Indonesia. Dengan demikian, penelitian ini berupaya tidak hanya menjawab tantangan akademis, tetapi juga memberikan manfaat praktis bagi peningkatan transparansi kebijakan dan penguatan kepercayaan masyarakat terhadap lembaga legislatif.

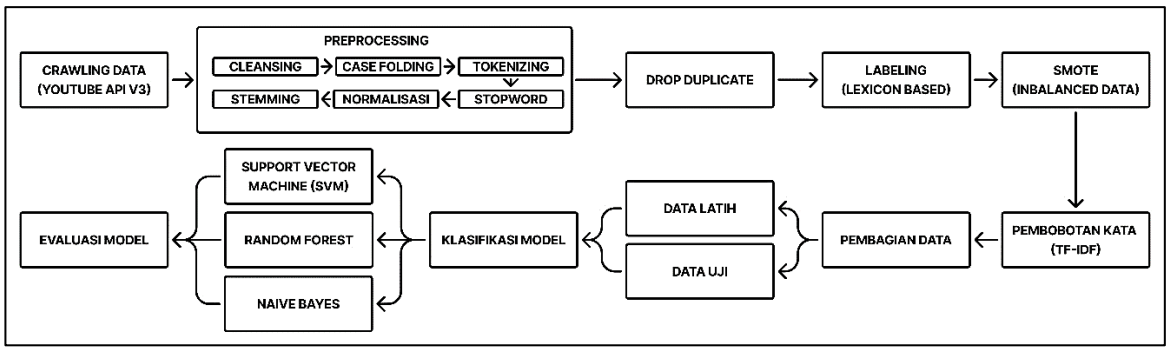
2. METODOLOGI PENELITIAN

2.1 Tahapan Penelitian

Penelitian ini dilaksanakan melalui tahapan yang sistematis guna memastikan hasil yang diperoleh valid dan dapat dipertanggungjawabkan. Setiap tahap dirancang untuk mendalami topik secara komprehensif serta memberikan

Building of Informatics, Technology and Science (BITS)
Volume 7, No 3, December 2025 Page: 1512−1524
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i3.8513

pemahaman yang lebih baik terhadap permasalahan yang dikaji[12]. Seluruh tahapan penelitian tersebut divisualisasikan dalam skema pada Gambar 1.



Gambar 1. Skema Penelitian

Berdasarkan Gambar 1 Alur penelitian ini dimulai dari pengambilan data komentar YouTube dengan YouTube API v3, kemudian dilakukan *preprocessing (cleansing, case folding, tokenizing, stopword removal, normalisasi, stemming)* hingga menghapus data duplikat. Data yang telah bersih diberi label sentimen berbasis *lexicon*, lalu diseimbangkan menggunakan SMOTE dan dibobotkan dengan TF-IDF. Selanjutnya data dibagi menjadi data latih dan data uji, kemudian diproses dengan algoritma *Naïve Bayes,* SVM, dan *Random Forest*. Hasil klasifikasi dievaluasi menggunakan metrik performa untuk mengetahui algoritma terbaik dalam analisis sentimen komentar YouTube terkait kenaikan tunjangan DPR RI.

2.2 Pengumpulan Data

Data diperoleh dari Komentar Penonton Melalui Youtube Data API Dengan Total Sebanyak 78.866 Baris Komentar Yang Dikumpulkan Dari Tujuh Video Yang Membahas Kontroversi DPR RI. Proses Pengambilan Data Dilakukan Menggunakan Modul *googleapiclient.discovery.build* Dengan API Versi V3, Di Mana Api_Key Berfungsi Sebagai Kunci *Autentikasi* Untuk Dapat Mengakses Data Dari Youtube. Informasi Mengenai Judul Video, Nama Kanal Youtube, Tanggal Unggah, Jumlah Tayangan, Serta Jumlah Komentar Dari Masing-Masing Video Disajikan Pada Tabel 1[13]. Seluruh Komentar Dikumpulkan Sejak Tanggal Unggah Masing-Masing Video, Sehingga Mencakup Seluruh Interaksi Penonton Yang Terekam Pada Platform Youtube. Adapun chanel youtube yang membahas topik ini bisa di lihat paada Tabel 1.

Tabel 1. *Channel Youtube*

No	Channel youtube	Judul vidio	Tanggal unggah	Jumlah tayangan	Jumlah komentar
1	Curhat Bang Denny Sumargo	Demo Runtuhkan Dpr, Pasha Ungu Menjawab Kenapa Dpr Berjoget	27 Agu 2025	730.316 x ditonton	28.268
2	Official iNews	Rakyat Bersuara Tunjangan DPR	3 Sep 2025	4.472.633 x ditonton	8.375
3	KompasTV Palu	DPR Akui Tunjangan & Gaji Ratusan Juta, Demo hingga Harapan Publik	29 Agu 2025	854.125 x ditonton	5.602
4	Kompas TV Jawa Timur	Keras! Diaspora Kritik Tunjangan Fantastis Anggota DPR, Soroti Pernyataan Ahmad Sahroni	27 Agu 2025	2.407.394 x ditonton	7.226
5	Kompas TV	residen Prabowo Bersama Megawati hingga Surya Paloh, Tanggapi Situasi usai Demo Kenaikan Tunajangan DPR	31 Agu 2025	4.027.062 x ditonton	19.959
6	Tribun news	Kediaman Anggota DPR Nafa Urbach Dijarah Massa, Warga: Rumah Porak-poranda Kini Sudah Dijaga Aparat akibat isu kenaikan tunjangan dpr	31 Agu 2025	3.117.827 x ditonton	7.066
7	CNN Indonesia	Geger Kabar Gaji Anggota DPR RI Naik Hingga Rp100 Juta	19 Agu 2025	224.737 x ditonton	2.370

Berdasarkan Tabel 1 yang menampilkan daftar video YouTube dari berbagai kanal berita dan individu yang membahas isu kenaikan tunjangan DPR RI pada Agustus–September 2025. Setiap video memiliki tingkat perhatian publik yang berbeda, terlihat dari jumlah tayangan dan komentar yang cukup tinggi, misalnya video dari *Official iNews* yang mencapai lebih dari 4,4 juta tayangan dengan 8.375 komentar, serta Kompas TV yang menembus 4 juta tayangan dengan hampir 20 ribu komentar. Hal ini menunjukkan bahwa topik kenaikan tunjangan DPR RI memicu diskusi luas dan respon signifikan dari masyarakat, baik dalam bentuk penayangan maupun komentar di ruang publik digital.

Building of Informatics, Technology and Science (BITS)
Volume 7, No 3, December 2025 Page: 1512−1524
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i3.8513

2.3 Preprocessing

Preprocessing data merupakan langkah awal dalam pengolahan data yang bertujuan untuk mempermudah analisis dan pemrosesan lebih lanjut. Proses ini mencakup pembersihan data dari gangguan, penyederhanaan dimensi data, serta penyusunan data agar lebih terstruktur. Tahap *preprocessing* penting untuk meningkatkan kualitas hasil analisis. Proses ini mencakup pembersihan data, normalisasi teks melalui *case folding*, pemecahan teks menjadi token (*tokenizing*), penghapusan kata tidak penting (*stopwords removal*), serta reduksi kata ke bentuk dasar (*stemming*)[14]. Berikut Penjelasannya:

a. *Cleansing* adalah proses membersihkan data dari noise seperti tanda baca dan karakter yang tidak relevan, Proses ini mencakup penghapusan elemen seperti indikator retweet (RT), hashtag (#), mention (@username), tautan (URL), simbol dan tanda baca seperti tanda seru (!) dan tanda tanya (?), angka, karakter non-alfabet, serta spasi berlebih.Langkah ini dilakukan untuk memastikan data lebih bersih, terstruktur, dan siap untuk analisis lebih lanjut[15].
b. *Case folding* adalah proses mengubah semua karakter dalam teks menjadi huruf kecil serta menghapus tanda bacadan angka. Proses ini hanya memproses huruf alfabet dari “a” hingga “z,” sehingga karakter selain huruf akan dihapus untuk memastikan data lebih konsisten dan siap untuk analisis[16].
c. *Tokenizing* adalah tahap di mana kalimat-kalimat dalam teks dipecah menjadi kata-kata tunggal atau tokenTujuan dari proses ini adalah untuk memberikan bobot nilai pada setiap dokumen dengan memecah karakterkarakter dalam teks menjadi satuan kata yang dapat dianalisis lebih lanjut [17].
d. *Stopword removal* adalah langkah untuk menghapus kata-kata yang tidak relevan, seperti "dan", "atau", dan "yang", karena tidak memberikan informasi penting [18].
e. *Normalisasi* adalah proses mengubah teks menjadi bentuk yang lebih terstruktur dan standar. Langkah inidilakukan dalam pemrosesan bahasa alami dan analisis teks untuk memastikan konsistensi dan keseragaman data.
f. *Stemming* adalah proses untuk memperoleh kata dasar dengan cara menghilangkan awalan, akhiran, kata sisipan, dan *confixes* (kombinasi awalan dan akhiran) menggunakan pustaka Sastrawi untuk kata-kata dalam bahasa Indonesia.

2.4 Drop Duplicate

Drop Duplicate adalah proses menghapus data yang memiliki nilai ganda atau berulang dalam sebuah dataset, sehingga hanya tersisa data unik yang relevan untuk dianalisis. Tahap ini penting dilakukan karena data duplikat dapat menyebabkan distorsi hasil analisis, menurunkan akurasi model, serta memperbesar ukuran data secara tidak efisien. Dengan menerapkan drop duplikat, kualitas data menjadi lebih bersih, terstruktur, dan representatif sehingga dapat meningkatkan keandalan hasil penelitian maupun kinerja algoritma analisis data[19].

2.4 Labeling

Tahap pelabelan menggunakan metode berbasis *lexicon* Bahasa Indonesia dengan memanfaatkan *InSet (Indonesian Sentiment Lexicon)* dilakukan dengan memanfaatkan kamus yang berisi katakata positif dan negatif. Proses ini didasarkan pada skor sentimen: jika skor lebih dari nol, maka sentimen dikategorikan sebagai positif; jika skor kurang nol, sentimen dikategorikan sebagai negatif. Sementara itu, jika skor tidak memenuhi kedua kriteria tersebut, sentimen akan dimasukkan ke dalam kategori netral[20]. Hal tersebut dapat dilihat pada Persamaan (1) berikut:

$$S_{sentiment} \begin{cases} Positif, & Jika\ Skor > 0 \\ Negatif, & Jika\ Skor < 0 \\ Netral, & Jika\ Skor = 0 \end{cases} \quad (1)$$

Persamaan (1) menunjukkan bahwa $S_{sentiment}$ merupakan label hasil klasifikasi sentimen berdasarkan skor *lexicon.* Skor dihitung dari penjumlahan nilai kata-kata pada komentar, di mana kata positif bernilai positif, kata negatif bernilai negatif, dan jika hasil penjumlahan bernilai nol maka komentar dikategorikan sebagai netral.

2.5 SMOTE

Proses pelabelan menghasilkan data yang tidak seimbang, sehingga penanganan lebih lanjut dibutuhkan menggunakan metode SMOTE (*Synthetic Minority Over-sampling Technique*). SMOTE adalah teknik pembelajaran mesin yang dirancang untuk mengatasi masalah ketidakseimbangan kelas dalam dataset. Teknik ini bekerja dengan membuat sampel baru dari kelas *minoritas* untuk menyeimbangkan dataset. Dengan menambahkan contoh dari kelas *minoritas*, dataset pelatihan menjadi lebih seimbang, dan data tambahan tersebut digunakan untuk melatih model pengklasifikasi[21].

2.6 Feature Extraction

TF-IDF adalah metode yang menggabungkan dua konsep, yaitu *term frequency* (TF) dan *inverse document frequency* (IDF). Metode ini digunakan untuk mengukur pentingnya suatu kata dalam sebuah data. TF menggambarkan seberapa sering suatu kata muncul dalam data tertentu, yang menandakan seberapa relevan kata tersebut dalam data tersebut[22]. Secara matematis algoritma ini dapat dituliskan seperti pada persamaan 2,3 dan 4.

Building of Informatics, Technology and Science (BITS)
Volume 7, No 3, December 2025 Page: 1512−1524
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i3.8513

$$tf = 0{,}5 + 0{,}5 \times \frac{tf}{\max(tf)} \quad (2)$$

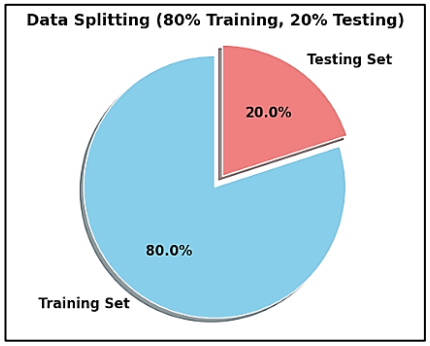
$$idf_t = log\left(\frac{D}{dft}\right) \quad (3)$$

$$W_{d.t} = tf_{d.t} \times idf_{d.t} \quad (4)$$

Persamaan 2,3 dan 4 merupakan perhitungan TF-IDF (*Term Frequency–Inverse Document Frequency*) yang digunakan untuk menentukan bobot suatu term dalam dokumen. Nilai *tf* menggambarkan jumlah kemunculan *term t* pada dokumen ke-*d*, sedangkan *idf* berfungsi mengukur tingkat kepentingan term dengan mempertimbangkan jumlah dokumen (*D*) yang mengandung term tersebut (*df*). Bobot akhir (*W*) dari *term* ke-*t* pada dokumen ke-*d* diperoleh dari hasil perkalian antara nilai *tf* dan *idf*, sehingga semakin sering suatu *term* muncul dalam dokumen tertentu namun jarang muncul di dokumen lain, maka bobotnya semakin tinggi. Pendekatan ini umum digunakan dalam analisis teks untuk mengidentifikasi kata-kata yang paling representatif dari suatu dokumen.

2.7 Data *Splitting*

Tahap berikutnya dalam penelitian ini adalah membagi data menjadi dua bagian, yaitu 80% sebagai data latih untuk mengembangkan model dan menyesuaikan parameter, serta 20% sebagai data uji untuk mengevaluasi kinerja model pada data yang belum pernah diproses[23]. Pembagian tersebut ditunjukkan pada Gambar 2.



Gambar 2. Pembagian Data Latih & Uji

2.8 Classification Naive Bayes Classifier

Naïve Bayes, yang didasarkan pada asumsi kuat tentang *independensi* setiap fitur atau kata dalam kelas yang diberikan. Algoritma *Naïve Bayes* bertujuan untuk memisahkan data ke dalam kelas-kelas yang spesifik. Kinerja pengklasifikasi dievaluasi berdasarkan akurasi prediksi. Algoritma ini memungkinkan sistem untuk secara akurat memahami dan memprediksi sentimen ulasan melalui analisis teks, menawarkan dasar yang kuat untuk menilai kualitas layanan aplikasi dan umpan balik pengguna. Metodologi penelitian melibatkan perincian masalah yang diteliti setepat mungkin (secara matematis, jika memungkinkan) dan menguraikan pendekatan yang diusulkan. Algoritma *Naïve Bayes* didasarkan pada *Teorema Bayes* dengan asumsi bahwa setiap fitur bersifat independen[24]. Rumus *Teorema Bayes* dapat ditulis Pada persamaan 5 sebagai berikut:

$$P(H \mid X) = \frac{P(H|X)P(H)}{P(X)} \quad (5)$$

Persamaan 5 merupakan sebuah rumus *Teorema Bayes* dimana rumus yang digunakan untuk menghitung *probabilitas* terjadinya suatu peristiwa berdasarkan informasi atau bukti yang ada. Dalam rumus tersebut, $(P(H|X))$ merepresentasikan *probabilitas* suatu *hipotesis* (H) benar setelah adanya bukti (X), dikenal sebagai *posterior probability*. *Probabilitas* ini dihitung dengan mengalikan $(P(X|H))$, yaitu peluang bukti (X) muncul jika *hipotesis* (H) benar (disebut *likelihood*), dengan $(P(H))$, yaitu *probabilitas* awal dari *hipotesis* (H) sebelum melihat bukti (disebut *prior probability*). Hasilnya kemudian dibagi dengan $(P(X))$, *probabilitas* keseluruhan dari bukti (X) terjadi tanpa memperhatikan *hipotesis* (H), yang dikenal sebagai marginal *likelihood*. Dengan *teorema* ini, kita dapat memperbarui keyakinan terhadap suatu *hipotesis* seiring dengan adanya bukti baru.

2.9 Classification Model Support Vector Mechine

Support Vector Machine (SVM) adalah algoritma *machine learning* yang menggunakan *hyperplane* sebagai pemisah antar kelas. Prediksi dilakukan dengan memberi label sesuai area kelas tempat data berada[25]. Rumus perhitungannya adalah sebagai berikut disajikan pada persamaan 6,7 dan 8.

$$\{(x_i, y_i)\}_{i=1}^{N} \quad (6)$$

Menghitung nilai w dan b:

Building of Informatics, Technology and Science (BITS)
Volume 7, No 3, December 2025 Page: 1512−1524
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i3.8513

$$w = \sum_{i=1}^{N} a_i\, , y_i, x_i$$

$$b = -\frac{1}{2}(w.x^{+} + w.x^{-}) \quad (7)$$

Fungsi Keputusan klasifikasi $sign\ (f(x))$

$$f(x) = w.x + b \text{ atau } f(x) = \sum_{i=1}^{m} a_i, y_i K(x, x_i) + b \quad (8)$$

N adalah jumlah data dalam dataset, n adalah jumlah fitur, dan m adalah jumlah *support vector* dengan $a_i > x$, yang menentukan margin optimal. $K(x, x_i)$ adalah fungsi kernel yang menghitung kedekatan titik data dalam ruang fitur lebih tinggi, memungkinkan SVM bekerja efektif tanpa menghitung koordinat eksplisit.

2.9 Classification Model Random Forest

Random Forest adalah metode *machine learning* yang dirancang untuk mengklasifikasikan dataset berukuran besar. Teknik ini merupakan pengembangan dari metode *Classification and Regression Tree* (CART). Dalam implementasinya, *Random Forest* menggunakan pendekatan *bootstrap aggregating* (bagging) dan secara acak memilih fitur saat membangun setiap pohon keputusan[26]. Hasil klasifikasi akhir diperoleh melalui voting *mayoritas* dari semua pohon yang dihasilkan. Dalam penelitian ini, *Random Forest* digunakan untuk memprediksi sentimen teks, apakah termasuk kategori positif, negatif, atau netral. Proses *decision tree* diawali dengan perhitungan nilai *gini impurity* dan rata-rata *gini impurity*. Rumus untuk menghitung *gini impurity* adalah sebagai berikut dapat dilihat pada persamaan 9,10 dan 11

$$Gini = 1 - \sum_{i}^{n} = 1\ (pi)^2 \quad (9)$$

dokumen n merujuk pada jumlah istilah dalam dokumen, sementara Pi adalah probabilitas kemunculan sebuah term, dihitung berdasarkan frekuensi kemunculannya dibandingkan dengan total term dalam dokumen.

$$Average\ gini\ impurity = \frac{n}{i} \times gini \quad (10)$$

i merujuk pada jumlah dokumen dalam dataset, sementara n adalah jumlah term dalam suatu kelas, yang bisa berupa kategori seperti positif, negatif, atau netral.

$$Information\ Gain\ =\ Gini\ impurity\ -\ average\ gini\ impurity \quad (11)$$

2.10 Evaluasi Model

Dalam klasifikasi multikelas, seperti Positive, Negatif, dan Netral, *confusion matrix* berbentuk tabel 3×3, di mana setiap baris mewakili kelas aktual dan setiap kolom menunjukkan kelas prediksi. Nilai diagonal menunjukkan jumlah data yang diklasifikasikan dengan benar (*True Positive*) untuk masing-masing kelas, sedangkan nilai di luar diagonal menunjukkan kesalahan prediksi (*False Positive* dan *False Negative*)[27]. Contoh confusion matrix untuk klasifikasi tiga kelas dapat dilihat pada Tabel 2 Sebagai berikut.

Tabel 2. *Confusion Matrix*

Actual	*Prediction*		
	Positive	*Negatif*	*Netral*
Positive	*True Positive* (TP)	*True Negative* (FN)	*False Negative* (FN)
Negatif	*False Positive* (FP)	*F True Positive* (TP)	*False Negative* (FN)
Netral	*False Positive* (FP)	*False Positive* (FP)	*True Positive* (TP)

Berdasarkan *confusion matrix* ini, metrik evaluasi seperti akurasi, presisi, recall, dan F1-Score dapat dihitung. Akurasi mengukur proporsi data yang terklasifikasi dengan benar terhadap keseluruhan data, presisi menilai ketepatan prediksi suatu kelas, recall menunjukkan kemampuan model mengenali seluruh data yang termasuk dalam kelas tersebut, dan F1-Score merupakan rata-rata harmonik dari presisi dan recall untuk menyeimbangkan keduanya. Untuk klasifikasi multikelas, metrik dihitung per kelas, kemudian dapat digabungkan menggunakan *macro average* (rata-rata sederhana dari semua kelas) atau *micro average* (perhitungan total TP, FP, FN seluruh kelas) untuk memperoleh evaluasi keseluruhan model. Rumus perhitungan metrik tersebut ditunjukkan pada Persamaan (12) hingga (15).

$$Akurasi = \frac{TP+TN}{TP+TN+FP+FN} \times 100\% \quad (12)$$

$$Presisi = \frac{TP}{TP+FP} \quad (13)$$

$$Recall = \frac{TP}{TP+FN} \quad (14)$$

$$f_1 - score = 2x = \frac{Presisi \times Recall}{Presisi + Recall} \quad (15)$$

Building of Informatics, Technology and Science (BITS)
Volume 7, No 3, December 2025 Page: 1512−1524
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i3.8513

3. HASIL DAN PEMBAHASAN

3.1 Data Collection

Data dikumpulkan melalui metode *crawling* dengan memanfaatkan modul *googleapiclient.discovery.build* menggunakan YouTube Data API v3 dan API Key sebagai kunci autentikasi untuk mengakses data dari *YouTube.* Proses ini berhasil memperoleh 78.866 baris komentar yang diambil dari tujuh video yang membahas kontroversi DPR RI Komentar tersebut dikumpulkan sejak tanggal unggah masing-masing video hingga data diambil, kemudian seluruh data disimpan dalam format CSV untuk diproses lebih lanjut pada tahap *preprocessing*. Dapat di lihat pada Tabel 3.

Tabel 3. Hasil Pengumpulan Data

No	*Tweet*
1	Di Indonesia itu pengenya gaji DPR disetarakan dgn parlement di eropa yg bergaji ($/euro) mangkannya hinggah seratus juta ketemunya. Padahal mereka ajah gak mampu untuk menurunkan nilai tukar ($/euro) ke Rp seperti zaman Pak Habibi. gini kok nyebut rakyat yg mau bubarin DPR itu t*l*l.Kalok gak mampu turunkan nilai tukar maunya bergaji setara dgn eropa tetap bubarin ajah.
78.866	itulah anggota dpr pintar bersilat lidah, gaji di ganti nama menjadi tunjangan... banyak rakyat yg sehari makan cuma 1x bahkan ada yg tidak makan seharian, ini malah wakil nya joget2 senang dpt tunjangan 100jt/ bulan.negeri apa coba ini... negara maju saja anggota dpr nya naik angkutan umum bukan naik alphard eh ini negara miskin anggota dpr nya ga ada yg naik angkutan umum...

Berdasarkan Tabel 2, hasil pengumpulan data menunjukkan jumlah komentar yang berhasil dikumpulkan. Langkah selanjutnya adalah melakukan *preprocessing* agar data lebih siap dan dinamis untuk dianalisis.

3.2 Preprocessing

Pada periode Agustus–September 2025, terkumpul 78.866 komentar YouTube yang kemudian diproses melalui tahapan *preprocessing*. Proses ini mencakup pembersihan teks dari karakter khusus, tanda baca, URL, *stopwords*, normalisasi kata, penghapusan duplikasi, serta tokenisasi Ringkasan hasil *preprocessing* ditampilkan pada Tabel 4.

Tabel 4. Hasil *Preprocessing*

Tahapan	*Tweet*
Data Tweet	Di Indonesia itu pengenya gaji DPR disetarakan dgn parlement di eropa yg bergaji ($/euro) mangkannya hinggah seratus juta ketemunya. Padahal mereka ajah gak mampu untuk menurunkan nilai tukar ($/euro) ke Rp seperti zaman Pak Habibi.
Data Cleaning	Di Indonesia itu pengenya gaji DPR disetarakan dgn parlement di eropa yg bergaji mangkannya hinggah seratus juta ketemunya Padahal mereka ajah gak mampu untuk menurunkan nilai tukar ke Rp seperti zaman Pak Habibi
Case Folding	di indonesia itu pengenya gaji dpr disetarakan dgn parlement di eropa yg bergaji mangkannya hinggah seratus juta ketemunya padahal mereka ajah gak mampu untuk menurunkan nilai tukar ke rp seperti zaman pak habibi
Normalize	[indonesia, ingin, gaji, dpr, disetarakan, dengan, parlemen, eropa, bergaji, hingga, seratus, juta, padahal, mereka, tidak, mampu, menurunkan, nilai, tukar, rupiah, seperti, zaman, habibi,]
Tokenizing	[‘indonesia’, ‘ingin’, ‘gaji’, ‘dpr’, ‘disetarakan’, ‘dengan’, ‘parlemen’, ‘eropa’, ‘bergaji’, ‘hingga’, ‘seratus’, ‘juta’, ‘padahal’, ‘mereka’, ‘tidak’, ‘mampu’, ‘menurunkan’, ‘nilai’, ‘tukar’, ‘rupiah’, ‘seperti’, ‘zaman’]
StopWords	[‘indonesia’, ‘gaji’, ‘dpr’, ‘disetarakan’, ‘parlemen’, ‘eropa’, ‘bergaji’, ‘seratus’, ‘juta’, ‘mereka’, ‘tidak’, ‘mampu’, ‘menurunkan’, ‘nilai’, ‘tukar’, ‘rupiah’]
Stemming	indonesia, gaji, dpr, setara, parlemen, eropa, gaji, seratus, juta, mereka, tidak, mampu, turun, nilai, tukar, rupiah.

Berdasarkan Tabel 3, *preprocessing* dilakukan untuk menyiapkan komentar agar lebih terstruktur. Tahapan meliputi data *cleaning* (menghapus simbol, emotikon, tanda baca), *case folding* (penyeragaman huruf kecil), *normalisasi* kata tidak baku, *tokenisasi, stopwords removal*, *serta stemming* untuk mengembalikan kata ke bentuk dasar. Hasil akhir menunjukkan teks lebih bersih, konsisten, dan siap digunakan dalam analisis sentimen.

3.3 Labeling

Setelah melalui tahap pemrosesan data, sebanyak 77.040 baris komentar berhasil disaring menjadi data yang bersih dan siap untuk dianalisis setelah di lakukan drop duplicate. Selanjutnya, setiap *tweet* dilabeli menggunakan *lexicon based* berdasarkan analisis sentimen Dapat dilihat pada Tabel 5 berikut untuk hasil pelabelan:

Building of Informatics, Technology and Science (BITS)
Volume 7, No 3, December 2025 Page: 1512−1524
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i3.8513

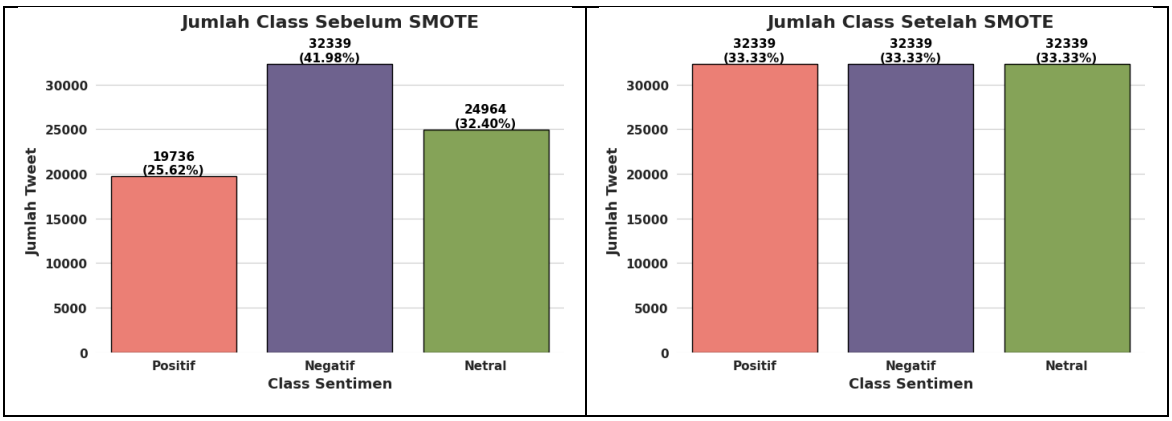
Tabel 5. Hasil Labeling

No	*Tweet*	*Sentimen Score*	*Label*
1	indonesia kenya gaji dpr tara parlement eropa gaji euro mang hinggah ratus juta ketemunyabrpadahal turun nilai tukar euro rp zaman habibi nyebut rakyat bubarin dpr tllbrkalok turun nilai tukar mau gaji tara eropa bubarin	4	Positif
2	mahasewa turun jalan rakyat turun jalan hadap polisi bayar pajak tahun jamaah perintah biar kelola sda bumn hasil negara	0	Netral
77.040	rakyat jerit kelaparandewan hormat mintak tunjangankami sesak napasmembunuh rakyat perlahanlebih kejam perang	-3	Negatif

Berdasarkan Tabel 4, proses pelabelan dilakukan dengan menghitung nilai polaritas pada rentang –1 hingga 1. Nilai polaritas yang lebih besar dari 0 dikategorikan sebagai sentimen positif, nilai yang lebih kecil dari 0 dikategorikan sebagai sentimen negatif, sedangkan nilai sama dengan 0 dikategorikan sebagai sentimen netral.

3.4 Handle Imbalance (SMOTE)

Analisis awal menunjukkan ketidakseimbangan kelas dengan 19.736 komentar positif, 32.339 negatif, dan 24.964 netral. Untuk mengatasinya digunakan *Synthetic Minority Oversampling Technique* (SMOTE), yang menambah sampel sintetis pada kelas minoritas sehingga distribusi data lebih seimbang dan model dapat mengenali pola sentimen secara lebih baik. Dapat dilihat pada Gambar 3.



Gambar 3. Perbandingan sentiment penggunaan Teknik SMOTE

Berdasarkan Gambar 3, terlihat perbandingan distribusi kelas sentimen sebelum dan sesudah penerapan SMOTE. Sebelum dilakukan SMOTE, jumlah data sentimen tidak seimbang, dengan 19.736 *tweet* berlabel positif (25,62%), 32.339 *tweet* negatif (41,98%), dan 24.964 *tweet* netral (32,40%). Ketidakseimbangan ini berpotensi menimbulkan bias pada model klasifikasi, karena model cenderung lebih akurat dalam memprediksi kelas *mayoritas* (negatif), namun kurang efektif dalam mengenali kelas *minoritas* (positif). Setelah penerapan SMOTE, distribusi data menjadi seimbang, di mana masing-masing kelas (positif, negatif, dan netral) memiliki jumlah tweet yang sama, yaitu 32.339 *tweet* (33,33%). Penyeimbangan ini sangat penting agar model dapat belajar secara proporsional dari setiap kelas, mengurangi bias terhadap kelas *mayoritas*, serta meningkatkan keandalan dan kinerja prediksi pada analisis sentimen secara keseluruhan.

3.5 Hasil Feature Extraction (TF-IDF)

TF-IDF *(Term Frequency - Inverse Document Frequency)* digunakan untuk mengukur tingkat kepentingan kata dalam dokumen dengan membandingkan frekuensi kemunculannya pada dokumen tertentu dan keseluruhan dokumen. Metode ini banyak diterapkan dalam analisis teks, pencarian informasi, dan pemrosesan bahasa alami (NLP).Hasil TF-IDF data dirangkum pada Tabel 6.

Tabel 6. Top 5 hasil TF-IDF

Term	TF_D1	TF_D2	TF_D3	IDF	TFIDF_D1	TFIDF_D2	TFIDF_D3
bicara	0	1	0	1,7	0	0,5	0
bubarin	0	1	1	1,7	0	0	0,27
dieorg	1	0	0	1,7	0,47	0	0
dpr	0	1	1	1,7	0	0	0,27
eropa	0	1	1	1,7	0	0	0,27

Berdasarkan Tabel 6 menampilkan hasil perhitungan TF-IDF *(Term Frequency – Inverse Document Frequency)* dari tiga dokumen (D1, D2, dan D3) terkait isu kenaikan gaji DPR RI. Nilai TF menunjukkan jumlah

Building of Informatics, Technology and Science (BITS)
Volume 7, No 3, December 2025 Page: 1512−1524
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i3.8513

kemunculan kata dalam setiap dokumen, sedangkan IDF mengukur tingkat kepentingannya berdasarkan distribusi kata di seluruh dokumen. Bobot TF-IDF, yaitu hasil kali TF dan IDF, menunjukkan relevansi kata terhadap dokumen tertentu. Kata dieorg memiliki bobot TF-IDF tertinggi (0,47 di D1), diikuti bicara (0,5 di D2) dan bubarin (0,27 di D3), sehingga dianggap paling signifikan dalam menggambarkan isi dokumen. Sebaliknya, kata seperti dpr dan eropa yang muncul di lebih banyak dokumen memiliki bobot TF-IDF rendah (0,27 di D3), sehingga kurang berpengaruh dalam membedakan isi antar dokumen. Hasil ini membantu mengidentifikasi kata kunci yang paling relevan untuk analisis sentimen pada topik yang dibahas.

3.6 Hasil Klasifikasi Model

Pada penelitian ini dilakukan perbandingan kinerja beberapa model klasifikasi sebelum dan sesudah penerapan teknik SMOTE (Synthetic Minority Oversampling Technique) sebagai upaya untuk mengatasi ketidakseimbangan kelas pada dataset. Tiga model yang diuji meliputi Naive Bayes, Support Vector Machine (SVM), dan Random Forest, yang masing-masing dianalisis untuk melihat perubahan performa setelah proses oversampling diterapkan. Evaluasi kinerja dilakukan menggunakan empat metrik utama, yaitu Accuracy, Precision, Recall, dan F1-Score, sehingga diperoleh gambaran yang lebih komprehensif mengenai efektivitas SMOTE dalam meningkatkan kemampuan model dalam mengenali kelas minoritas. Hasil perbandingan ini diharapkan dapat menunjukkan model mana yang paling optimal dan seberapa besar kontribusi SMOTE dalam meningkatkan kualitas prediksi.Hasil klasifikiasi disajikan pada Tabel 7 dan Tabel 8.

Tabel 7. Hasil Klasifikasi *Before* SMOTE

Model	*Accurasy*	*Class*	*Precision*	*Recall*	*F1-Score*
Nave Bayes	62%	Negatif	56%	97%	72%
		Netral	69%	32%	44%
		Positif	91%	43%	59%
Support Vector Mechine	89%	Negatif	95%	90%	93%
		Netral	81%	88%	84%
		Positif	90%	88%	89%
Random Forest	81%	Negatif	79%	92%	85%
		Netral	83%	66%	74%
		Positif	81%	79%	80%

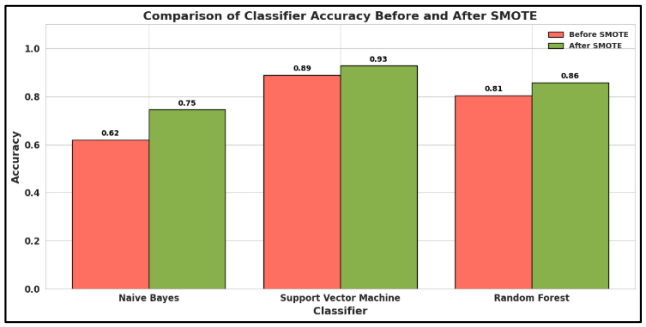
Berdasarkan Tabel 7 yang menunjukkan hasil klasifikasi sebelum penerapan SMOTE, akurasi model yang diperoleh adalah *Naive Bayes* sebesar 62%, *Support Vector Machine* (SVM) sebesar 89%, dan *Random Forest* sebesar 81%. Hasil klasifikasi memperlihatkan adanya ketidakseimbangan performa antar kelas. *Naive Bayes* mencatatkan kinerja terendah, terutama pada kelas netral dengan *precision* 69%, *recall* 32%, dan *F1-score* 44%, yang menandakan kesulitan dalam mendeteksi kelas ini. Sementara itu, SVM tampil paling konsisten dengan akurasi tinggi, serta nilai *precision*, *recall*, dan *F1-score* yang seimbang di semua kelas (misalnya pada kelas positif dengan *F1-score* 89%). *Random Forest* menempati posisi menengah dengan akurasi 81%, menunjukkan performa cukup baik pada kelas negatif (*F1-score* 85%), namun masih kurang optimal dalam mengenali kelas netral (*F1-score* 74%). Secara keseluruhan, SVM terbukti menjadi model dengan performa terbaik sebelum penerapan SMOTE.

Tabel 8. Hasil Klasifikasi *After* SMOTE

Model	*Accurasy*	*Class*	*Precision*	*Recall*	*F1-Score*
Nave Bayes	75%	Negatif	71%	88%	79%
		Netral	80%	47%	59%
		Positif	76%	88%	81%
Support Vector Mechine	93%	Negatif	95%	91%	93%
		Netral	88%	92%	90%
		Positif	96%	95%	95%
Random Forest	86%	Negatif	84%	89%	86%
		Netral	87%	76%	81%
		Positif	86%	93%	89%

Berdasarkan Tabel 8 yang menunjukkan hasil setelah penerapan SMOTE, terlihat adanya peningkatan kinerja model dalam mendeteksi setiap kelas. Pada *Naive Bayes*, akurasi mencapai 75% dengan nilai *Precision* 76% dan *Recall* 88% pada kelas positif, menghasilkan *F1-Score* sebesar 81%. Model *Support Vector Machine* (SVM) menunjukkan kinerja paling unggul dengan akurasi tertinggi 93%, serta *Precision* 96% dan *Recall* 95% pada kelas positif dengan *F1-Score* sebesar 95%. Sementara itu, *Random Forest* memperoleh akurasi 86% dengan *Precision* 86% dan *Recall* 93% pada kelas positif sehingga menghasilkan *F1-Score* sebesar 89%. Secara keseluruhan, penerapan SMOTE membantu meningkatkan kemampuan semua model, dengan SVM sebagai model terbaik dalam mendeteksi kelas positif dan menangani ketidakseimbangan data. Untuk memberikan pemahaman yang lebih jelas mengenai hasil akurasi, perbandingan ketiga model sebelum dan sesudah penerapan SMOTE ditampilkan pada Gambar 4.

Building of Informatics, Technology and Science (BITS)
Volume 7, No 3, December 2025 Page: 1512−1524
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i3.8513

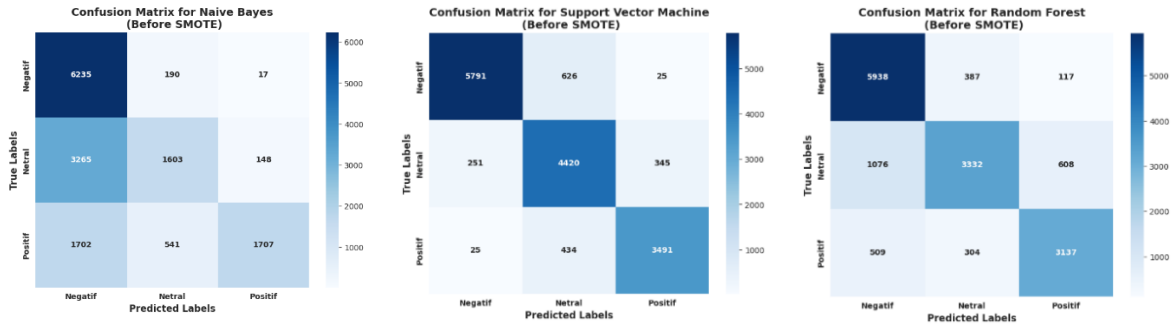


Gambar 5. *Perbandingan Klasifikasi*

Gambar 5 menunjukkan perbandingan akurasi klasifikasi sebelum dan sesudah penerapan SMOTE, yang secara konsisten meningkatkan performa ketiga metode pembelajaran mesin. *Naïve Bayes* mengalami peningkatan akurasi dari 62% menjadi 75%, SVM dari 89% menjadi 93%, dan *Random Forest* dari 81% menjadi 86%. Hasil ini menegaskan bahwa SMOTE efektif dalam menyeimbangkan data sehingga model dapat belajar lebih optimal, dengan SVM tetap mencatat akurasi tertinggi baik sebelum maupun sesudah penerapan SMOTE.

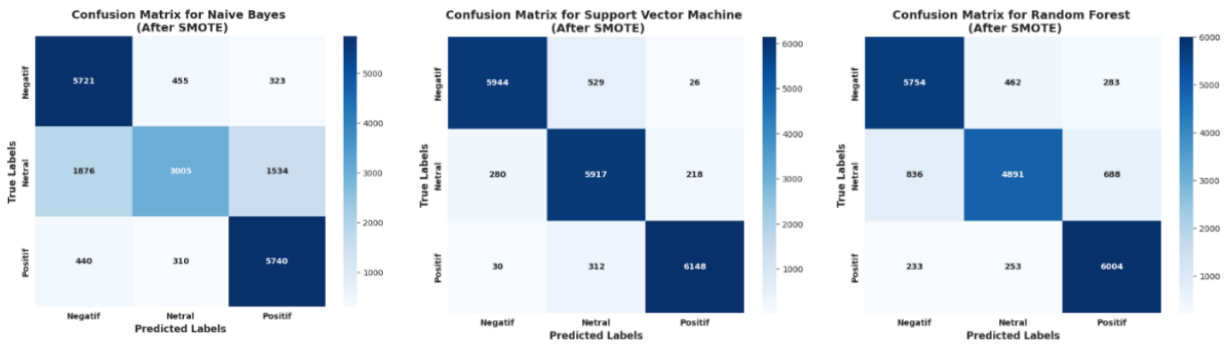
3.7 Evaluasi Model

Setelah melakukan klasifikasi dengan *Naive Bayes*, *Support Vector Machine* (SVM), dan *Random Forest*, evaluasi kinerja model dilakukan menggunakan *Confusion Matrix*. Evaluasi kinerja model dilakukan menggunakan *Confusion Matrix* untuk mengukur *True Positive* (TP), *True Negative* (TN), *False Positive* (FP), dan *False Negative* (FN), yang digunakan untuk menghitung metrik seperti *Precision, Recall, dan F1-Score*. Dapat dilihat pada gambar 6 dan 7.



Gambar 6. *Confusion Matrix Before* SMOTE

Berdasarkan Gambar 6 yang menunjukkan hasil evaluasi menggunakan *confusion matrix* sebelum penerapan SMOTE, algoritma *Support Vector Machine* (SVM) memiliki *True Positive* (TP) sebanyak 3.491, *True Negative* (TN) sebanyak 5.791, *False Positive* (FP) sebanyak 25, dan *False Negative* (FN) sebanyak 434, yang menandakan kinerja cukup baik namun masih lemah dalam mendeteksi kelas netral. *Random Forest* menghasilkan TP sebanyak 3.137, TN sebanyak 5.938, FP sebanyak 117, dan FN sebanyak 304, dengan prediksi yang relatif stabil meskipun kesalahan pada kelas netral cukup tinggi. Sementara itu, *Naïve Bayes* memperoleh TP sebanyak 1.707, TN sebanyak 6.235, FP sebanyak 17, dan FN sebanyak 541, menunjukkan kesulitan signifikan dalam mengklasifikasikan kelas positif dan netral dibandingkan dua algoritma lainnya.



Gambar 7. *Confusion Matrix After* SMOTE

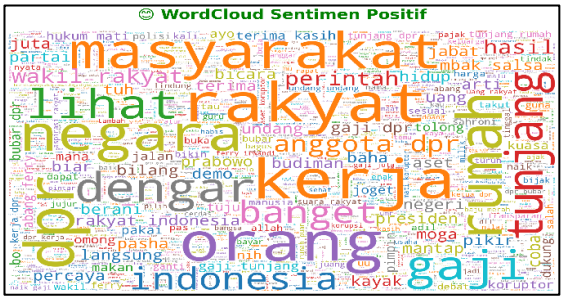
Berdasarkan Gambar 7 yang menunjukkan hasil evaluasi menggunakan confusion matrix setelah penerapan SMOTE, algoritma *Support Vector Machine* (SVM) memiliki *True Positive* (TP) sebanyak 6.148, *True Negative* (TN)

Building of Informatics, Technology and Science (BITS)
Volume 7, No 3, December 2025 Page: 1512−1524
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i3.8513

sebanyak 5.944, *False Positive* (FP) sebanyak 26, dan *False Negative* (FN) sebanyak 312, menandakan peningkatan signifikan dalam mendeteksi kelas positif dan netral. *Random Forest* menghasilkan TP sebanyak 6.004, TN sebanyak 5.754, FP sebanyak 283, dan FN sebanyak 253, dengan prediksi yang lebih seimbang di seluruh kelas. Sementara itu, *Naïve Bayes* memperoleh TP sebanyak 5.740, TN sebanyak 5.721, FP sebanyak 323, dan FN sebanyak 310, menunjukkan peningkatan kinerja yang cukup baik meskipun masih terjadi kesalahan lebih tinggi dibandingkan SVM dan *Random Forest*.

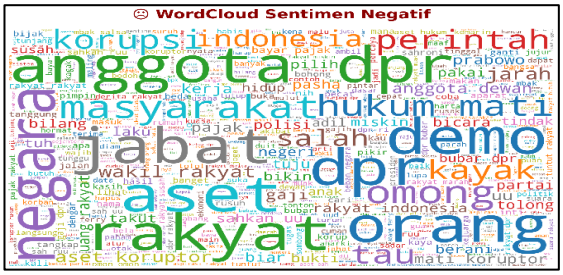
3.8 Visualisasi WordCloud dan Frekuensi Kata

Visualisasi *WordCloud* digunakan untuk menampilkan kata-kata yang paling sering muncul pada sentimen positif, negatif, dan netral terkait tunjangan DPR RI, dengan ukuran kata mencerminkan tingkat frekuensinya. Dapat dilihat pada Gambar 8,9 dan Gambar 10.



Gambar 8. *Wordcloud* Positif

Gambar 8 menunjukkan *WordCloud* sentimen positif menampilkan kata dominan seperti masyarakat, kerja, lihat, dengar, negara, rakyat, terima kasih, bagus, percaya, dan mantap. Kata-kata ini menunjukkan adanya apresiasi, dukungan, serta harapan optimis dari masyarakat terhadap pemerintah maupun DPR. Sentimen positif sering muncul dalam bentuk pujian terhadap kerja, ungkapan rasa syukur, ajakan untuk percaya pada institusi, dan harapan baik bagi masa depan bangsa. Kehadiran kata seperti terima kasih, bagus, mantap, dan moga memperkuat nuansa optimis dan semangat kolaboratif dalam percakapan publik.



Gambar 9. *Wordcloud* Negatif

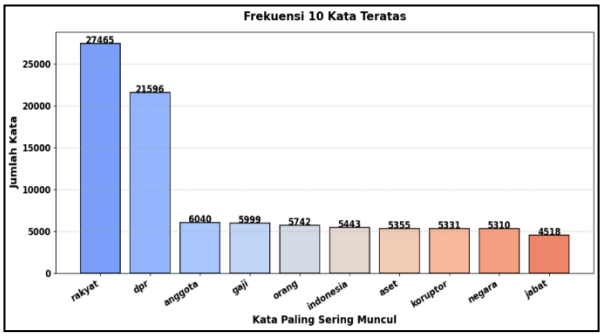
Gambar 9 menunjukkan *WordCloud* sentimen negatif memperlihatkan dominasi kata seperti anggota DPR, korupsi, jabatan, demo, salah, omong, hukum, mati, dan rakyat. Kata-kata tersebut menunjukkan bahwa komentar dengan sentimen negatif dipenuhi kritik keras terhadap DPR maupun pemerintah, terutama terkait isu korupsi, penyalahgunaan jabatan, dan kebijakan yang dianggap merugikan rakyat. Nada percakapan cenderung emosional, terlihat dari kata-kata seperti salah, mati, atau demo, yang mencerminkan kekecewaan, kemarahan, bahkan seruan aksi. Hal ini menggambarkan bahwa sentimen negatif digunakan untuk mengekspresikan ketidakpuasan dan tuntutan perubahan.

Gambar 10. *Wordcloud* Netral

Gambar 10 menunjukkan *WordCloud* sentimen netral menunjukkan bahwa percakapan publik banyak berkisar pada kata orang, masyarakat, Indonesia, rakyat, pemerintah, hukum, negara, dan DPR. Kata-kata ini menandakan

Building of Informatics, Technology and Science (BITS)
Volume 7, No 3, December 2025 Page: 1512−1524
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i3.8513

bahwa komentar bernuansa netral lebih sering membicarakan isu-isu umum terkait pemerintahan, fungsi DPR, aturan hukum, serta peran masyarakat tanpa disertai emosi yang berlebihan. *Netralitas* ini biasanya muncul ketika pengguna sekadar menyampaikan informasi, opini faktual, atau membicarakan topik politik dan hukum secara deskriptif tanpa memberikan penilaian baik maupun buruk.



Gambar 11. Frekuensi kata

Berdasarkan Gambar 11 yang menunjukan grafik frekuensi 10 kata teratas, terlihat bahwa kata "rakyat" (27.465 kali) dan "dpr" (21.596 kali) mendominasi pembahasan, menandakan bahwa isu terkait hubungan antara DPR dan rakyat menjadi topik yang paling banyak diperbincangkan. Kata lain seperti "anggota," "gaji," "orang," dan "indonesia" juga sering muncul, menunjukkan fokus pembicaraan pada individu maupun lembaga dalam konteks kebijakan dan representasi rakyat. Sementara itu, kata "aset," "koruptor," "negara," dan "jabat" memperlihatkan adanya perhatian terhadap isu korupsi, pengelolaan aset negara, serta jabatan politik. Secara keseluruhan, grafik ini menggambarkan bahwa perbincangan publik lebih banyak berpusat pada kritik dan sorotan terhadap DPR serta perannya dalam mengelola kepentingan rakyat dan negara.

4. KESIMPULAN

Berdasarkan hasil penelitian, algoritma *Support Vector Machine*, *Naïve Bayes*, dan *Random Forest* menunjukkan perbedaan kinerja dalam analisis sentimen komentar YouTube terkait pembahasan kenaikan tunjangan DPR RI. Sebelum penerapan SMOTE, SVM memperoleh akurasi 89%, *Random Forest* 81%, dan *Naïve Bayes* 62%, dengan kelemahan utama pada klasifikasi sentimen netral. Setelah penerapan SMOTE, seluruh algoritma mengalami peningkatan, dengan SVM mencapai akurasi tertinggi sebesar 93%, sehingga terbukti menjadi algoritma paling unggul dalam penelitian ini. Visualisasi *WordCloud* menunjukkan kata dominan berbeda pada tiap kategori sentimen. Sentimen positif didominasi kata seperti masyarakat, kerja, rakyat, dan terima kasih yang merefleksikan apresiasi dan harapan optimis terhadap pemerintah dan DPR. Sentimen negatif menonjolkan kata seperti anggota DPR, korupsi, jabatan, dan demo yang menggambarkan kritik keras, kekecewaan, serta ketidakpuasan publik terhadap kebijakan tersebut. Sementara itu, sentimen netral memuat kata seperti orang, masyarakat, dan negara yang bersifat deskriptif tanpa emosi kuat. Secara umum, sentimen negatif lebih mendominasi, mencerminkan persepsi publik yang kritis terhadap DPR RI. Penelitian ini juga membuktikan bahwa penerapan SMOTE efektif meningkatkan kinerja model, dan disarankan penelitian selanjutnya membandingkan metode *balancing* lain untuk hasil yang lebih optimal.

REFERENCES

[1] A. D. Putra, A. S. R. Wulandari, And A. Arief, "Indonesian Political Dynamics In National And Regional Elections," *Amsir Law Journal*, Vol. 6, No. 1, Pp. 28–39, Oct. 2024, Doi: 10.36746/Alj.V6i1.595.

[2] A. Setiawan, S. M. Sari, And N. Rahmah, "Lembaga Negara Pemegang Kekuasaan Legislatif," *Jurnal Hukum Dan Kewarganegaraan*, Vol. 15, 2025, Doi: 10.3783/Causa.V2i9.2461.

[3] J. Hukum And C. Justitia, "Analisis Hak Recall Partai Politik Terhadap Anggota Dewan Perwakilan Rakyat," *Jurnal Hukum Caraka Justitia*, Vol. 5, No. 1, Pp. 43–54, 2025, Doi: Https://Doi.Org/10.30588/Jhcj.V5i1.2192.

[4] Anak Teknik Indonesia, "Isu Kenaikan Gaji Dpr Ri Tahun 2025: Klarifikasi, Kontroversi, Dan Tuntutan Transparansi," Https://Www.Anakteknik.Co.Id/. Accessed: Sep. 24, 2025. [Online]. Available: Https://Www.Anakteknik.Co.Id/Araa_Fii/Articles/Isu-Kenaikan-Gaji-Dpr-Ri-Tahun-2025-Klarifikasi-Kontroversi-Dan-Tuntutan-Transparansi?Srsltid=Afmboordzr15ukgntmbejokqdsbeiqn-Xytrdkmzhwdu0lblaalp01qh&

[5] B. Wicaksono And V. R. S. Nastiti, "Analisis Sentimen Dalam Opini Publik Di Chanel Youtube Indonesia Lawyers Club Tentang Isu Populer Dengan Menggunakan Metode Lstm Dan Bi-Lstm," *Jurnal Algoritma*, Vol. 21, No. 2, Pp. 241–251, Dec. 2024, Doi: 10.33364/Algoritma/V.21-2.1696.

[6] A. Sudiongko, "8 Negara Dengan Jumlah Pemakai Youtube Terbanyak Di Dunia," Madiun.Jatimtimes.Com. Accessed: Sep. 24, 2025. [Online]. Available: Https://Madiun.Jatimtimes.Com/Baca/284336/20230224/072500/8-Negara-Dengan-Jumlah-Pemakai-Youtube-Terbanyak-Di-Dunia

[7] I. Koswara, "Global Komunika Diskursus Digital Dan Suara Publik: Komentar Youtube Dalam Kasus #Indonesiagelap," *Global Komunika*, Vol. 8, No. 2, 2025.

Building of Informatics, Technology and Science (BITS)
Volume 7, No 3, December 2025 Page: 1512−1524
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i3.8513

[8] M. Bais, A. Hakiki, And Y. Darmi, "Penerapan Algoritma Machine Learning Svm Dan Nbc Pada Sentimen Analisis Komentar Youtube Program Pengaduan Masyarakat Lapor Mas Wapres," *Jim: Jurnal Multidisiplin*, Vol. 4, No. 1, 2025, Doi: 10.38035/Jim.V4i1.

[9] T. Putri, Siti Nurhaliza, And Dodi Vionanda, "Analisis Sentimen Penggunaan Aplikasi Youtube Menggunakan Metode Naïve Bayes," *Unp Journal Of Statistics And Data Science*, Vol. 3, No. 1, Pp. 60–66, Feb. 2025, Doi: 10.24036/Ujsds/Vol3-Iss1/343.

[11] A. R. Andriawan, M. Mustakim, And R. Novita, "Sentiment Analysis Classification Of Political Parties On Twitter Using Gated Recurrent Unit Algorithm And Natural Language Processing," *Journal Of Informatics And Telecommunication Engineering*, Vol. 7, No. 2, Pp. 514–522, Jan. 2024, Doi: 10.31289/Jite.V7i2.10709.

[12] M. Jamil, H. Hadiyanto, And R. Sanjaya, "Sentiment Analysis: Classifying Public Comments On Youtube In Disaster Management Simulation In Indonesia Using Naïve Bayes And Support Vector Machine," *Ingenierie Des Systemes D'information*, Vol. 29, No. 2, Pp. 437–446, Apr. 2024, Doi: 10.18280/Isi.290205.

[13] U. Krishna, "Youtube Comments Sentiments Analysis," *Int J Res Appl Sci Eng Technol*, Vol. 13, No. 1, Pp. 875–880, Jan. 2025, Doi: 10.22214/Ijraset.2025.66475.

[14] S. K. Dirjen, P. Riset, D. Pengembangan, R. Dikti, S. Khomsah, And A. S. Aribowo, "Terakreditasi Sinta Peringkat 2 Model Text-Preprocessing Komentar Youtube Dalam Bahasa Indonesia," *Iaii*, Vol. 1, No. 3, Pp. 648–654, 2017, Doi: Https://Doi.Org/10.29207/Resti.V4i4.2035.

[15] H. Ali And N. Hendrastuty, "Comparison Of Naïve Bayes Classifier, Support Vector Machine, Random Forest Algorithms For Public Sentiment Analysis Of Kip-K Program On Twitter," *Jurnal Teknik Informatika (Jutif)*, Vol. 5, No. 6, Pp. 1701–1712, Dec. 2024, Doi: 10.52436/1.Jutif.2024.5.6.4030.

[16] F. Aldi Wijaya And Parjito, "Analisis Sentimen Publik Terhadap Virus Hmpv Berdasarkan Media Sosial X Dengan Algoritma Logistic Regression," *Building Of Informatics, Technology And Science (Bits)*, 2025, Doi: 10.47065/Bits.V9i9.999.

[17] S. Al Afif And R. R. Suryono, "Analisis Sentiment Publik Terhadap Deepfake Ai Menggunakan Aplikasi X Dengan Metode Support Vector Mechine Dan Naive Bayes Classifier," *Building Of Informatics, Technology And Science (Bits)*, 2025, Doi: 10.47065/Bits.V9i9.999.

[18] L. Kencono And D. Darwis, "Perbandingan Algoritma Nbc, Svm Dan Random Forest Untuk Analisis Sentimen Implementasi Starlink Pada Media Sosial X," *Technology And Science (Bits)*, Vol. 6, No. 4, 2025, Doi: 10.47065/Bits.V6i4.6813.

[19] S. F. Huwaida, R. Kusumawati, And B. Isnaini, "Analisis Sentimen Komentar Youtube Terhadap Pemindahan Ibu Kota Negara Menggunakan Metode Naïve Bayes," *Jambura Journal Of Informatics*, Vol. 6, No. 1, Pp. 26–39, Apr. 2024, Doi: 10.37905/Jji.V6i1.24718.

[20] M. Fernanda And N. Fathoni, "Perbandingan Performa Labeling Lexicon Inset Dan Vader Pada Analisa Sentimen Rohingya Di Aplikasi X Dengan Svm," *Jurnal Informatika Dan Sains Teknologi*, Vol. 1, No. 3, Pp. 62–76, 2024, Doi: 10.62951/Modem.V1i3.112.

[21] T. Chamidy, "Application Of Smote In Sentiment Analysis Of Myxl User Reviews On Google Play Store," *Jurnal Informatika Sunan Kalijaga)*, Vol. 10, No. 1, Pp. 74–86, 2025, Doi: Https://Doi.Org/10.14421/Jiska.2025.10.1.74-86.

[22] D. Sugiarto, E. Utami, And A. Yaqin, "Perbandingan Kinerja Model Tf-Idf Dan Bow Untuk Klasifikasi Opini Publik Tentang Kebijakan Blt Minyak Goreng," *Jti*, Dec. 2022, Doi: Https://Doi.Org/10.25105/Jti.V12i3.15669.

[23] T. Arya Bimantoro, Y. Rahmawati, Y. Findawati, And M. A. Rosid, "Sentiment Analysis Of Youtube Comments On The East Java Grant Case Using The Support Vector Machine (Svm) Algorithm. [Analisis Sentimen Komentar Youtube Terhadap Kasus Dana Hibah Jawa Timur Menggunakan Algoritma Support Vector Machine (Svm)]," *Department Of Informatics Engineering*, 2025, Doi: Https://Doi.Org/10.21070/Ups.9030.

[24] F. Fatma Wati, A. Eko Widodo, And P. Korespondensi, "Analisis Sentimen Ulasan Pengguna Aplikasi Deepseek Menggunakan Algoritma Random Forest Dan Naive Bayes," *Computer And Network Technology*, Vol. 5, No. 1, Pp. 8–15, 2025, Doi: Https://Doi.Org/10.31294/Hqpha267.

[25] M. Bais, A. Hakiki, And Y. Darmi, "Penerapan Algoritma Machine Learning Svm Dan Nbc Pada Sentimen Analisis Komentar Youtube Program Pengaduan Masyarakat Lapor Mas Wapres," *Jim: Jurnal Ilmu Multidisiplin*, Vol. 4, No. 1, 2025, Doi: 10.38035/Jim.V4i1.

[26] T. Cahya Herdiyani And A. U. Zailani, "Sentiment Analysis Terkait Pemindahan Ibu Kota Indonesia Menggunakan Metode Random Forest Berdasarkan Tweet Warga Negara Indonesia Sentiment Analysis Related To Transportation Of Indonesian Capital City Using Random Forest Method Based On Tweet Of Indonesian Citizens," *Jtsi*, Vol. 3, No. 2, Pp. 154–165, 2022, Doi: Https://Doi.Org/10.35957/Jtsi.V3i2.2920.

[27] F. Panjaitan, W. Ce, H. Oktafiandi, G. Kanugrahan, Y. Ramdhani, And V. H. C. Putra, "Evaluation Of Machine Learning Models For Sentiment Analysis In The South Sumatra Governor Election Using Data Balancing Techniques," *Journal Of Information Systems And Informatics*, Vol. 7, No. 1, Pp. 461–478, Mar. 2025, Doi: 10.51519/Journalisi.V7i1.1019