JUMANJI | Vol. 8 | No. 1 | Halaman 1 - 15
ISSN (p): 2598-8050 April 2024
ISSN (e): 2598-8069

Penerapan Metode Support Vector Machine dan SMOTE untuk Klasifikasi Sentimen Publik Terhadap Polisi Republik Indonesia

Fitri Destiyanti[1], Asep Id Hadiana[2], Fajri Rakhmat Umbara[3]

[1,2,3] Jurusan Teknik Informatika, Fakultas Sains dan Informatika
Universitas Jenderal Achmad Yani
Jl. Terusan Jenderal Jenderal Sudirman, Cimahi
Email : fitridestiyan29@gmail.com

ABSTRAK

Analisis sentimen atau ekstraksi opini adalah penelitian yang mengevaluasi sudut pandang, pemikiran, serta persepsi mengenai berbagai topik, subjek, dan produk dengan memanfaatkan data opini yang tersedia pada platform media sosial.. Platform media sosial populer seperti YouTube, khususnya melalui saluran "86 & Custom Protection NET" yang bekerjasama dengan Kepolisian Negara Republik Indonesia, menyajikan aktivitas polisi dan mendapat respons dari masyarakat dalam bentuk komentar. Komentar-komentar ini menjadi sumber data dalam penelitian text mining untuk mengklasifikasikan sentimen positif atau negatif. Penelitian ini menggunakan pendekatan menggunakan machine learning dengan metode Support Vector Machine (SVM) dan teknik SMOTE untuk menangani ketidakseimbangan data dalam komentar youtube. Hasil analisis menunjukkan akurasi sebesar 91%, dengan presisi 63%, recall 68%, dan f1 score 65% berdasarkan perhitungan confusion matrix.

***Kata kunci**: Media Sosial, Support Vector Machine, Text Mining, Youtube.*

ABSTRACT

Sentiment analysis or opinion extraction is a research method that evaluates perspectives, thoughts, and perceptions on various topics, subjects, and products by utilizing opinion data available on social media platforms. Popular social media platforms such as YouTube, particularly through the channel "86 & Custom Protection NET" in collaboration with the Indonesian National Police, present police activities and receive responses from the public in the form of comments. These comments serve as data sources in text mining research to classify positive or negative sentiments. This study employs a machine learning approach using the Support Vector Machine (SVM) method and the SMOTE technique to address data imbalance in YouTube comments. The analysis results show an accuracy of 91%, with precision of 63%, recall of 68%, and an F1 score of 65% based on the confusion matrix calculation.

***Keywords**: Social Media, Support Vector Machine, Text Mining, Youtube.*

1. PENDAHULUAN

Kepolisian Negara Republik Indonesia memiliki tanggung jawab untuk memastikan keselamatan dan ketertiban umum, melaksanakan penegakan hukum sesuai dengan peraturan yang berlaku, serta memberikan perlindungan, bantuan, dan layanan kepada warga masyarakat guna menciptakan keamanan di dalam negeri sesuai dengan ketentuan hukum yang berlaku (Rahmawati and Silvi n.d.). Pembentukan Polri dilakukan dengan tujuan untuk menjaga terciptanya kedamaian dalam masyarakat yang menghormati hak-hak asasi manusia (Zuber and Program Doktor Ilmu Hukum 2017). Kepolisian mendapatkan perhatian dari masyarakat terkait pelaksanaan tugas, tanggung jawab, dan kinerjanya.

Berbagai isu seperti kasus pembunuhan, pelecehan seksual, tindakan korupsi dan kasus lainnya yang terjadi dalam setahun terakhir telah mengakibatkan penurunan kepercayaan masyarakat terhadap kinerja kepolisian. Kepercayaan masyarakat menurun karena tugas polisi dalam memberikan keamanan dan pengayoman serta pelayanan yang masih kurang dan jauh dari harapan masyarakat. Pada masa kini, di dalam era digital, kemajuan teknologi informasi dan internet, khususnya dalam konteks media social, telah mengalami pertumbuhan yang sangat cepat., memberikan kemudahan bagi masyarakat dalam mengakses informasi dengan cepat. Hashtag #PercumaLaporPolisi pernah menjadi trending topik pada media sosial yang menjadi tempat keluhan dan kritik terkait kasus-kasus dan aduan masyarakat terhadap kepolisian yang bertindak tidak profesional dalam menangani laporan masyarakat (Hidayatullah, Alam, and Jaelani n.d.)

Youtube merupakan salah satu platform yang popular dengan jumlah pengguna yang sangat besar (Munthe et al. n.d.). Saluran YouTube 86 & Custom Protection NET adalah sebuah program realitas yang bekerja sama dengan Kepolisian Negara Republik Indonesia (Polri) untuk menayangkan kegiatan dan kinerja polisi dalam menjalankan tugasnya mulai dari mendisiplinkan pengguna dan menjaga ketertiban lalu lintas, penangkapan pelaku kriminal, penggerebekan, penangkapan sindikat narkoba sampai kasus berat kepolisian. Masyarakat banyak memberikan feedback atau respon mengenai kepolisian dari tayangan tersebut, opini yang diberikan beragam dan dapat bersifat negatif, positif atau netral. Komentar tersebut memiliki potensi untuk menjadi sumber data yang berguna dalam konteks penelitian. Analisis sentimen merupakan salah satu cara penggunaan dari teknik text mining., di mana sentimen umumnya merujuk kepada perasaan, pendapat, sikap, atau opini yang terkait dengan subjek tertentu, seperti individu, komunitas, entitas, peristiwa, atau topik. (Amrullah et al. 2020).

Dataset yang tidak seimbang atau *imbalanced* adalah kondisi yang seringkali ditemui dalam pengumpulan data langsung. Dataset yang tidak seimbang terjadi ketika distribusi sampel antara kelas-kelas yang berbeda dalam dataset tidak seimbang. Dengan kata lain, ada satu atau beberapa kelas yang memiliki jumlah sampel yang jauh lebih rendah daripada kelas-kelas lainnya. Kondisi *imabalanced* dataset ini dapat mengakibatkan kinerja model klasifikasi menjadi kurang optimal. Ini disebabkan oleh fakta bahwa algoritma cenderung memberikan label pada kelas mayoritas saat melakukan prediksi, mengabaikan kelas minoritas. Akibatnya, kelas mayoritas cenderung menunjukkan tingkat akurasi yang lebih tinggi. (Dwi Fitriani et al. 2021). Penelitian sebelumnya pada tahun 2022 yang dilakukan oleh Magnolia (Magnolia et al. 2022) telah melakukan studi dalam menangani *imbalanced* dataset, khususnya dalam studi kasus komentar Twitter terhadap program Kampus Merdeka di tingkat Perguruan Tinggi. Dalam penelitian tersebut, metode Support Vector Machine digunakan untuk menangani masalah ketidakseimbangan data ini. Melalui penerapan teknik oversampling ADASYN dengan variasi nilai *max_features* yang berbeda, penelitian tersebut berhasil meningkatkan tingkat akurasi.

Algoritma Support Vector Machine (SVM) merupakan metode dalam *machine learning* yang digunakan untuk melakukan klasifikasi dengan tujuan mencari *hyperplane* optimal yang dapat memisahkan dua kelas yang berbeda dalam ruang input.(Analisis Dan Penerapan et al. 2019). SMOTE merupakan teknik oversampling yang digunakan untuk mengembangkan dataset pada kelas minoritas dengan menciptakan data tambahan melalui reproduksi data yang ada dalam kelas minoritas tersebut. Dalam penggunaan metode SMOTE, peningkatan jumlah sampel pada kelas minoritas dilakukan dengan cara mengambil sampel dari kelas tersebut dan mencari k-nearest neighbor dari setiap sampel yang diambil. Selanjutnya, contoh sintetis dihasilkan dengan cara merangkai contoh minoritas yang ada, pendekatan ini membantu menghindari terjadinya *overfitting* yang berlebihan (Sutoyo et al. n.d.)

Dalam penelitian ini, dilakukan analisis komentar-komentar di YouTube terkait polisi dengan menggunakan metode klasifikasi Support Vector Machine (SVM) dan menerapkan teknik SMOTE untuk mengatasi ketidakseimbangan dataset pada data yang digunakan. Data berasal dari komentar-komentar tersebut yang kemudian dikelompokkan menjadi sentimen positif dan negatif. Pemberian bobot pada kata-kata dalam dokumen dilakukan dengan metode *Term Frequency-Inverse Document Frequency* (TF-IDF). Evaluasi performa model dalam mengklasifikasikan komentar-komentar di YouTube dilakukan dengan menggunakan *Confusion Matrix.*

2. METODE

2.1. Pengumpulan Data

Data yang digunakan dalam penelitian ini merupakan data yang diperoleh dari hasil proses *crawling* komentar menggunakan Bahasa pemrograman python pada media social youtube. Data diambil dari komentar youtube pada tayangan di *channel* 86 & Custom Protection NET, dimulai dari tanggal 15 Desember sampai 30 April 2023. Dari hasil *crawling*, data yang diambil sebanyak 1000 dataset.

2.2. *Preprocessing*

Langkah awal dalam proses data mining adalah tahap *preprocessing*, yang bertujuan untuk mengubah data asli yang diperoleh dari YouTube API menjadi data yang siap digunakan dalam langkah selanjutnya. Proses ini mencakup upaya menghilangkan atau mengurangi gangguan (*noise*) sehingga data dapat diubah menjadi informasi yang lebih terperinci dan efisien untuk pengolahan lanjutan. Proses yang dilakukan dalam tahapan *preprocessing* data diantaranya:

a. *Cleaning*

Cleaning atau proses pembersihan, pada tahapan ini bertujuan untuk menghilangkan karakter yang tidak relevan atau noise pada data teks yang akan diproses. Noise tersebut dapat berupa tanda baca, karakter khusus, tag HTML, URL dan lainnya. Tujuan proses cleaning untuk mempersiapkan teks yang lebih bersih dan terstruktur untuk tahapan selanjutnya.

b. *Case folding*

Proses *Case folding* merupakan suatu tindakan yang diterapkan dengan tujuan untuk merubah seluruh teks yang terdapat dalam dokumen ke dalam suatu format yang seragam, sehingga memastikan konsistensi dalam penulisan karakter huruf besar dan huruf kecil. Sebagai ilustrasi, dalam kasus ini, kata "DATA" dan "data" akan dianggap

sebagai hal yang berbeda oleh sistem, oleh karena itu perlu dilakukan penyesuaian terhadap teks untuk membuatnya memiliki bentuk yang sama. Dalam penelitian ini, seluruh teks di dalam dokumen akan diubah menjadi huruf kecil atau *lowercase*.

c. *Tokenizing*

Tokenisasi adalah tahap dimana teks dipecah menjadi unit-unit lebih kecil yang dikenal sebagai token, yang dapat berupa segmen huruf, kata, atau bahkan kalimat, sebelum dilakukan analisis lebih lanjut. Token bisa mencakup berbagai jenis entitas seperti kata, angka, simbol, tanda baca, dan lainnya.

d. *Stemming*

Stemming adalah Langkah atau hahapan dalam preprocessing yang dimana teks diubah menjadi bentuk kata dasarnya dengan cara menghapus imbuhan-imbuhan yang ada pada teks tersebut.

e. *Stopward removal*

Dalam tahap eliminasi *stopword*, digunakan untuk secara selektif menghapus kata-kata yang secara konsisten muncul dalam teks dokumen dan secara substansial tidak memberikan kontribusi signifikan terhadap pemahaman konten yang relevan. Ini mencakup kata-kata umum seperti "di, yang, atau, dan," serta kata-kata serupa lainnya.

f. *Replace slang word*

Replace slang word adalah langkah yang digunakan untuk mengganti kata-kata slang dengan bentuk kata standar atau baku.

2.3. *Lexicon Based Features*

Salah satu teknik yang dapat digunakan untuk mengevaluasi apakah suatu kalimat, teks, atau komentar mengungkapkan sentimen yang bersifat positif, negatif, atau netral adalah dengan menggunakan metode berdasarkan leksikon. Pendekatan ini melibatkan identifikasi kata-kata dan frasa yang memiliki makna positif, negatif, atau netral, serta melakukan analisis terhadap pola penggunaan kata-kata tersebut dalam teks tersebut (Syakur 2021) Fitur berbasis leksikon atau *Lexicon Based Features* adalah jenis fitur yang berasal dari pengetahuan dan berfokus pada ekstraksi leksikon berdasarkan pendapat yang terdapat dalam teks, dengan tujuan mengidentifikasi polaritas dari leksikon tersebut. Leksikon adalah kumpulan istilah yang telah dikenali. Fitur ini memberikan nilai atau bobot berdasarkan penggunaan leksikon atau kamus untuk mengategorikan dokumen sebagai sentimen positif atau negatif (Desai and Mehta 2016).

2.4. *Term Frequency – Inverse Document Frequency (TF-IDF)*

Perhitungan TF-IDF melibatkan penggabungan persamaan *Term Frequency* (TF) dan persamaan *Inverse Document Frequency* (IDF). Dalam perhitungan *Term Frequency* (TF), mengevaluasi seberapa sering suatu fitur (*term*) muncul dalam teks, dengan nilai TF yang semakin tinggi seiring dengan frekuensi munculnya yang lebih besar. Sebaliknya, dalam persamaan *Inverse Document Frequency* (IDF), nilai IDF suatu fitur akan menurun jika fitur tersebut cenderung muncul dalam banyak teks, sementara nilai IDF akan meningkat jika fitur tersebut hanya muncul dalam sedikit teks. Pendekatan TF-IDF ini berguna untuk menilai

signifikansi sebuah kata dalam suatu dokumen dan memberikan bobot pada hubungan antara kata tersebut dan dokumen tersebut (Jurnal, Maulana, and Witanti 2023)

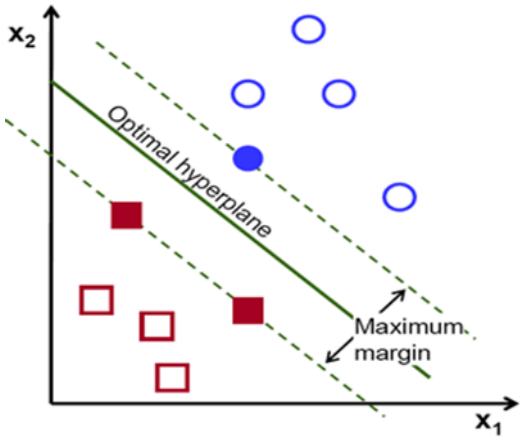
Metode perhitungan Term Frequency-Inverse Document Frequency (TF-IDF) yang digunakan dalam penelitian menggunakan persamaan (1)

$$TF_{IDF_{(ij)}} = \mathrm{tf_{ij}} * (\log\left(\frac{N}{n_i}\right) + 1) \quad (1)$$

Dalam rumus ini, tf_{ij} menggambarkan frekuensi kemunculan fitur i dalam teks j, N adalah total jumlah teks, dan n_i adalah jumlah teks yang mengandung fitur i.

2.4. Support Vector Machine

Teknik Support Vector Machine (SVM) merupakan sebuah pendekatan relatif baru dalam dunia analisis data, baik untuk tugas klasifikasi maupun regresi. SVM tergolong dalam kategori pembelajaran terawasi (*supervised learning*) yang melibatkan dua tahap utama, yakni fase pelatihan dengan metode sequential training SVM, yang diikuti oleh tahap pengujian. Pada dasarnya, inti dari penggunaan SVM dalam klasifikasi adalah mencari *hyperplane* optimal yang bertindak sebagai pemisah antara dua kelompok data. Keunggulan SVM terletak pada kemampuannya untuk mengatasi dataset yang memiliki dimensi yang tinggi, berkat teknik kernel yang digunakan. Dalam algoritma SVM, hanya sejumlah titik data yang dipilih secara khusus, dikenal sebagai support vector, yang memegang peran sentral dalam membentuk model yang digunakan dalam proses klasifikasi (Athira Luqyana, Cholissodin, and Perdana 2018). Penggambaran metode Support Vector Machine dapat dilihat pada Gambar di bawah ini.



Gambar 1. Ilustrasi Support Vector Machine

2.5. SMOTE *(Synthetic Minority Over-sampling Technique)*

Salah satu teknik *oversampling* yaitu SMOTE, merupakan metode yang digunakan untuk meningkatkan jumlah data pada kelas minoritas. Metode ini dilakukan dengan menciptakan data sintetis yang didasarkan pada data yang sudah ada di kelas minoritas. Dalam pelaksanaan SMOTE, proses *oversampling* mengambil sampel dari kelas minoritas, mengidentifikasi tetangga terdekat (k-nearest neighbor) untuk setiap sampel tersebut, dan menghasilkan sampel-sampel sintetis sebagai opsi tambahan daripada sekadar menggandakan sampel-

sampel asli dari kelas minoritas. Dengan pendekatan ini, SMOTE membantu mengatasi masalah *overfitting* yang mungkin terjadi (Sutoyo et al. n.d.)

2.4. Evaluasi *Confusion Matrix*

Confusion Matrix adalah salah satu alat khusus yang digunakan untuk menilai kinerja dalam situasi yang dapat diklasifikasikan yang melibatkan dua nilai atau lebih. *Confusion Matrix* memberikan informasi tentang True Positive (kasus yang teridentifikasi jelas positif), True Negative (kasus yang jelas teridentifikasi negatif), False Positive (kasus yang teridentifikasi jelas positif),yang teridentifikasi jelas positif), dan False Negative (kasus yang teridentifikasi jelas negatif) dalam bentuk tabel yang mencakup empat kombinasi berbeda yaitu hasil prediksi dan hasil nilai sebenarnya. *Confusion Matrix* memberikan wawasan penting tentang kualitas prediksi model klasifikasi, membantu dalam mengevaluasi performa dan mengidentifikasi tingkat kesalahan dan keberhasilan dalam mengklasifikasikan data pada setiap kelas (Deng et al. 2016). Dapat dilihat pada Tabel 1. *Confusion Matrix*.

Tabel 1. Confusion Matrix

Nilai sebenarnya	Nilai prediksi	
	True	False
True	True Positive (TP)	False Positive (FP)
False	False Negative (FN)	True Negative (TN)

Oleh karena itu untuk mengukur performa dari algoritma *Support Vector Machine* ini menggunakan empat metode evaluasi pada *Confusion Matrix*. Metode tersebut yaitu :

a. *Accuracy* adalah sebuah metrik evaluasi yang digunakan dalam pemodelan klasifikasi untuk menilai tingkat keberhasilan model klasifikasi dalam memperkirakan kelas dengan benar dari data yang diamati. Akurasi menghitung persentase prediksi yang tepat (true positives dan true negatives) dibandingkan dengan jumlah total data yang diamati.

b. *Precision* adalah ukuran yang menilai sejauh mana prediksi yang bernilai positif yang diberikan oleh model benar-benar akurat. Ini merupakan perbandingan antara true positives dengan jumlah keseluruhan antara true positives dan false positives.

c. *Recall* adalah metrik yang mengukur sejauh mana kemampuan model dalam mengenali semua kasus positif yang sebenarnya. Ini didefinisikan sebagai perbandingan antara jumlah true positives dengan jumlah true positives dan false negatives.

d. *F1-score,* metrik evaluasi yang digunakan dalam pemodelan klasifikasi untuk mengukur seimbang antara precision (presisi) dan recall (recall) dalam kinerja model. Ini memberikan gambaran keseluruhan tentang sejauh mana model tersebut memiliki keseimbangan yang baik antara kemampuan untuk memberikan prediksi yang benar

(precision) dan kemampuan untuk mengidentifikasi sebanyak mungkin kasus positif yang sebenarnya (recall).

3. HASIL DAN PEMBAHASAN

3.1. Pengumpulan Data

Pada penelitian yang dilakukan, data yang digunakan adalah hasil data diperoleh dari komentar pengguna youtube pada tayangan channel 86 & Custom Protection NET. Data komentar yang digunakan untuk penelitian diperoleh dengan melakukan *crawling* dengan memanfaatkan API Youtube dan Bahasa pemrograman python. Data komentar yang diambil yaitu sebanyak 1000 komentar youtube yang dimulai dari15 Desember 2022 hingga 30 April 2023. Contoh data yang telah diambil dapat dilihat dalam Tabel 2. Contoh Pengumpulan Data di bawah ini.

Tabel 2. Contoh Pengumpulan Data Komentar

Username	*Comment*	*Like Count*	*Publish Date*
MAS MV	Kenapa polisi banyak yang galak dan tegas padahal saya gak takut	1	21-03-2023
Sudarmanto Bekonang	Terima kasih Pak Polisi. Semoga ini menjadi kegiatan rutin dilakukan disemua daerah daerah di negeri ini dan tambah berkah team polri.	10	28-03-2023

3.2. Data Pre-processing

Preprocessing adalah langkah pertama dalam proses data mining, dilakukan dengan cara mengubah data *raw* atau mentah yang telah dikumpulkan dari youtube ke dalam format yang siap digunakan untuk tahapan berikutnya. Tujuannya adalah mengurangi atau menghilangkan gangguan (*noise*) agar data menjadi lebih bersih dan siap digunakan dalam proses lebih lanjut. Pada penelitian ini, terdapat lima langkah dalam tahap praproses yang dilakukan, yakni membersihkan data, mengubah huruf menjadi huruf kecil, membagi data menjadi token, menormalkan kata, menghapus kata-kata pengisi (*stopword*), dan menggantikan kata-kata *slang.* Hasil dataset yang telah melalui tahap *preprocessing* ditunjukkan pada Tabel 3.

Tabel 3. Data Hasil Preprocessing

Username	*Comment*	*Like Count*	*Publish Date*
Anto Pratama	saya malam habis kena keroyok pas gilir ketangakep polsek polres diem pada nyali seperti gitu saja sama polisi saja sudah takut kamu	0	01-02-2023

	bikin sampah masyarakat saja kamu nyusahin orang tua		
tarek hamood	polisi lembek banget hormat anak anak nakal kaya gini	0	09-03-2023

3.3. Pelabelan Data

Pemberian label pada data merupakan langkah untuk mengklasifikasikan atau memberikan kelas pada data komentar mentah agar dapat digunakan dalam tahap selanjutnya. Pemberian label pada data melibatkan pengelompokan ke kategori positif atau negatif. Metodenya berdasarkan jumlah kata positif dan negatif dalam kalimat. Jika kata positif lebih banyak, kalimat diklasifikasikan sebagai positif, sebaliknya untuk negatif. Kamus yang dipakai adalah "Indonesia Sentimen (InSet) Lexicon," yakni sebuah kamus leksikon Bahasa Indonesia yang disusun oleh Fajri Koto dan Gemal Y. Rahmaningtyas. Pemilihan kamus InSet dilakukan karena kamus ini memberikan tingkat akurasi yang optimal dalam konteks leksikon Bahasa Indonesia. InSet terdiri dari 3.609 kata dengan makna positif dan 6.609 kata dengan makna negatif. Jika jumlah kata yang menunjukkan respons positif lebih banyak daripada jumlah kata yang menunjukkan respons negatif, maka kalimat tersebut akan diklasifikasikan sebagai positif, dan sebaliknya. Setiap kata yang menunjukkan respons positif akan diberi nilai +1, sedangkan kata yang menunjukkan respons negatif akan diberi nilai -1.Jika skor akhir adalah 0, maka kalimat tersebut diklasifikasikan sebagai netral. Oleh karena itu, diperlukan pengecekan manual untuk benar-benar mengklasifikasikan data menjadi positif atau negatif. Proses dimulai dengan mengimpor data yang telah dibersihkan. Kemudian, dilakukan pembacaan kata-kata positif dan negatif. Daftar kata-kata yang menunjukkan respons positif diperoleh dari file "positive.txt", sedangkan kata-kata yang menunjukkan respons negatif diperoleh dari file "negative.txt". Setelah itu, label diberikan pada setiap dokumen.

Tabel 4. Pelabelan Data

Username	Comment	Score_pos	Score_neg	Label
MAS MV	polisi galak gak takut	0	-4	Negatif
cakra ningrat	dukung ambarita pangakat sekolah promosi rakyat butuh polisi jujur	7	-6	Positif

3.4. Pembobotan TF-IDF

Setelah tahapan *preprocessing* selesai dilakukan, selanjutnya proses pembobotan dokumen menggunakan TF-IDF. Data yang akan dijadikan bahan dalam penerapan TF-IDF telah

menjalani proses *preprocessing* sebelumnya. Pada tabel dibawah ini terdapat contoh penggunaan delapan data dengan rincian sebagai berikut:

Tabel 5. Contoh Dataset

No.	Teks	Data
1.	apresiasi kerja keras polisi republik indonesia berantas tindak jahat semangat	Train Positif 1
2.	terima kasih polisi ku cinta	Train Positif 2
3.	terima kasih polisi kerja bagus	Train Positif 3
4.	bikin cacat celaka polisi sampah masyarakat kapok tawuran resah masyarakat	Train Negatif 1
5.	aneh polisi pakai obat terlarang	Train Negatif 2
6.	polisi baju merah tidak sopan bahasa coba kuruptor siksa	Train Negatif 3
7.	semangat polisi mantap masyarakat percaya	Test Positif
8.	polisi arogan benar anggap salah	Test Negatif

Dalam proses pembobotan menggunakan TF-IDF, terdapat tiga tahapan: pertama, menciptakan vektor berdasarkan TF-IDF; kedua, mendapatkan nilai vektor berdasarkan IDF; dan terakhir, mengalikan hasil perhitungan TF dengan IDF. Metode TF-IDF digunakan untuk menghitung bobot kata-kata yang umum digunakan dan dikenal karena efisiensinya serta kemudahan implementasinya. Dalam metode ini, TF dan IDF dihitung untuk setiap kata dalam setiap dokumen dalam korpus untuk menentukan seberapa sering kata tertentu muncul dalam dokumen. Pada tabel dibawah ini merupakan hasil proses pembobotan dokumen menggunakan *Term Frequency – Inverse Document Frequency (TF-IDF) sebagai berikut.*

Tabel 6. Hasil TF-IDF

DF	IDF	TF-IDF							
		Train P1	Train P2	Train P3	Train N1	Train N2	Train N3	Test P	Test N
1	0.90308999	0.90309	0	0	0	0	0	0	0
2	0.60205999	0.60206	0	0.60206	0	0	0	0	0
1	0.90308999	0.90309	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0
1	0.90308999	0.90309	0	0	0	0	0	0	0

DF	IDF	TF-IDF							
		Train P1	Train P2	Train P3	Train N1	Train N2	Train N3	Test P	Test N
1	0.90308999	0.90309	0	0	0	0	0	0	0
1	0.90308999	0.90309	0	0	0	0	0	0	0
1	0.90308999	0.90309	0	0	0	0	0	0	0
2	0.60205999	0.60206	0	0	0	0	0	0.60206	0
2	0.60205999	0	0.60206	0.60206	0	0	0	0	0
2	0.60205999	0	0.60206	0.60206	0	0	0	0	0
1	0.90308999	0	0.90309	0	0	0	0	0	0
1	0.90308999	0	0.90309	0	0	0	0	0	0
1	0.90308999	0	0	0	0.90309	0	0	0	0
1	0.90308999	0	0	0	0.90309	0	0	0	0
1	0.90308999	0	0	0	0.90309	0	0	0	0
1	0.90308999	0	0	0	0.90309	0	0	0	0
2	0.60205999	0	0	0	0.60206	0	0	0.60206	0
1	0.90308999	0	0	0	0.90309	0	0	0	0
1	0.90308999	0	0	0	0.90309	0	0	0	0
1	0.90308999	0	0	0	0.90309	0	0	0	0
1	0.90308999	0	0	0	0	0.90309	0	0	0
1	0.90308999	0	0	0	0	0.90309	0	0	0
1	0.90308999	0	0	0	0	0.90309	0	0	0
1	0.90308999	0	0	0	0	0.90309	0	0	0
1	0.90308999	0	0	0	0	0	0.90309	0	0
1	0.90308999	0	0	0	0	0	0.90309	0	0
1	0.90308999	0	0	0	0	0	0.90309	0	0
1	0.90308999	0	0	0	0	0	0.90309	0	0
1	0.90308999	0	0	0	0	0	0.90309	0	0
1	0.90308999	0	0	0	0	0	0.90309	0	0

DF	IDF	TF-IDF							
		Train P1	Train P2	Train P3	Train N1	Train N2	Train N3	Test P	Test N
1	0.90308999	0	0	0	0	0	0.90309	0	0
1	0.90308999	0	0	0	0	0	0.90309	0	0
1	0.90308999	0	0	0	0	0	0	0.90309	0
1	0.90308999	0	0	0	0	0	0	0.90309	0
1	0.90308999	0	0	0	0	0	0	0	0.90309
1	0.90308999	0	0	0	0	0	0	0	0.90309
1	0.90308999	0	0	0	0	0	0	0	0.90309
1	0.90308999	0	0	0	0	0	0	0	0.90309
1	0.90308999	0.90309	0	0	0	0	0	0	0

3.5. Klasifikasi Support Vector Machine Berbasis SMOTE

Setelah menyelesaikan tahap pembobotan atau ekstraksi fitur dengan metode TF-IDF, langkah berikutnya adalah melaksanakan proses klasifikasi dengan menggunakan algoritma Support Vector Machine (SVM). SVM adalah suatu metode pembelajaran mesin yang sering digunakan dalam permasalahan regresi dan klasifikasi, terutama dalam konteks klasifikasi. Karena data yang digunakan dari hasil *crawling* komentar youtube tidak seimbang, dilakukan pendekatan normalisasi sampling dilakukan melalui metode *oversampling* dengan *Synthetic Minority Oversampling Technique* (SMOTE) untuk menyeimbangkan distribusi jumlah data dengan meningkatkan jumlah data pada kelas minoritas. *Oversampling* adalah strategi yang digunakan untuk menyamakan proporsi jumlah data dengan meningkatkan volume data pada kelas yang kurang dominan. Pendekatan ini dimaksudkan untuk menaikkan jumlah data. Teknik *oversampling* SMOTE, dalam hal ini, berfungsi dengan cara menciptakan data sintetis tambahan dari kelas yang kurang dominan. Setelah proses penyeimbangan data dengan menggunakan SMOTE selesai,, langkah berikutnya adalah mengembangkan model klasifikasi menggunakan SVM. SVM menemukan garis pemisah atau yang disebut sebagai *hyperplane* terbaik antara dua kelas data, kemudian meningkatkan akurasi dengan memproyeksinya ke dalam ruang dimensi yang lebih tinggi. Tujuannya adalah agar model dapat membedakan dan mengklasifikasikan sentimen dengan lebih akurat dan generalisasi yang lebih baik.

3.5. Eksperimen dan Pengujian

Pengujian yang dilakukan melibatkan perbandingan akurasi dua metode Support Vector Machine, yakni satu tanpa menggunakan SMOTE dan satu pengujian dengan penerapan SMOTE. Pengujian ini dilakukan dengan membandingkan empat perbandingan rasio pengambilan data latih dan data uji yang berbeda, yakni rasio 70:30, 75:25, 80:20 dan 90:10 seperti yang tercatat dalam Tabel 7. Berikut adalah hasil dari pengujian tersebut

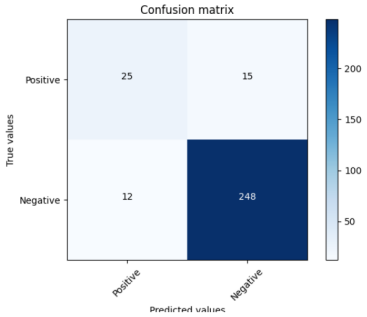
Tabel 7. Hasil Pengujian

Rasio 70:30 (700)		Rasio 75:25 (750)		Rasio 80:20 (800)		Rasio 90:10 (900)	
SVM	SVM Berbasis SMOTE	SVM	SVM Berbasis SMOTE	SVM	SVM Berbasis SMOTE	SVM	SVM Berbasis SMOTE
84.67%	*91%*	*85.20%*	*88.40%*	*84.50%*	*86%*	*85%*	*90%*

Pada Tabel 7 dari hasil eksperimen yang dilakukan pada pengujian menggunakan SVM memberikan hasil untuk rasio 70:30 menghasilkan nilai akurasi 84,67%, rasio 75:25 menghasilkan rata-rata 85,20%, rasio 80:20 menghasilkan rata-rata 84,50%, rasio 90:10 menghasilkan rata-rata 85%. Hasil pengujian SVM berbasis SMOTE pada rasio 70:30 menghasilkan rata-rata 91%, rasio 75:25 menghasilkan rata-rata 88,40%, rasio 80:20 menghasilkan rata-rata 86% dan pada rasio 90:10 menghasilkan rata-rata 90%.

Tujuan pengujian ini adalah untuk mengevaluasi tingkat akurasi dalam mengklasifikasikan komentar youtube menggunakan sistem yang telah dikembangkan. Sistem yang diuji terdiri dari dua sentimen, yaitu positif dan negatif. Untuk menguji akurasi, digunakan Confusion Matrix yang membandingkan hasil prediksi dengan kelas asli dari data input. Dari hasil eksperimen pengujian yang dilakukan, maka pengambilan rasio terbaik yang diterapkan yaitu rasio 70:30 yaitu menggunakan data latih 30% dari total 1000 data yang digunakan, dan mendapatkan akurasi paling tinggi yaitu 91% dengan kombinasi SMOTE pada klasifikasi menggunakan Support Vector Machine.

Pengujian hasil prediksi yang dilakukan oleh Support Vector Machine dan SMOTE pada data Uji menggunakan metode confusion matrix dapat dilihat pada Gambar 2.



Gambar 2. Hasil Confusion Matrix

Berikut adalah perhitungan *Confusion Matrix* dengan cara menghitung nilai *Accuracy, Precision, Recall dan F1-Score* sebagai berikut :

a. Precision

Precision adalah ukuran seberapa akurat model memprediksi data positif. Nilainya diperoleh dengan menghitung hasil bagi antara jumlah data positif yang terklasifikasi

secara benar sebagai positif dan jumlah data positif yang salah terklasifikasi sebagai positif, termasuk data positif yang seharusnya diklasifikasikan sebagai negatif.

Precision : TP / (TP + FP)

- *TP : True Positif*
- *FP : False Positif*

= 25/(25+15)
= 25/40
= 0,625 x 100%
= 62,5 % (dibulatkan jadi 63%)

b. Recall

Recall adalah ukuran seberapa lengkap model dalam memprediksi data positif. Nilainya dihitung dengan membagi jumlah data positif yang benar-benar positif dengan jumlah total data positif, termasuk data positif yang sebenarnya negatif.

(1) *Recall* = TP / (TP +FN)

- *TP : True Positif*
- *FN : False Negatif*

= 25/(25+12)
= 25/37
= 0,6756 x 100%
= 67,56 % (dibulatkan jadi 68%)

c. *F1-Score* atau *F-Measurement*

F-measure adalah satu parameter tunggal yang mengukur seberapa berhasil sistem pengambilan informasi dengan menggabungkan nilai recall dan precision. Nilai *F-measure* dihitung dengan mengalikan hasil dari precision dan recall, kemudian membaginya dengan hasil penjumlahan precision dan recall, lalu dikalikan dengan dua.

(2) *F-Measure*=2*(*Precision*Recall)/ (Precision+Recall)*

= 2*(63*68) / (63+68)
= 2*(4.284)/(131)
= 8.568 / 131
= 65,40 x 100%
= 65 %

d. Akurasi

Akurasi adalah metrik yang menunjukkan sejauh mana model mampu mengidentifikasi sentimen secara akurat. Ini dihitung dengan membagi jumlah sentimen yang benar-benar teridentifikasi dengan benar oleh model dengan jumlah total data, termasuk data uji.
Rumus :

$$Akurasi = \frac{TP + TN}{TP + TN + FP + FN} \text{ X } 100\% \qquad (2)$$

Hasil perhitungan :

$$Akurasi = \frac{25 + 248}{25 + 248 + 15 + 12} \text{ X } 100\%$$

$$Akurasi = 0{,}91\ x\ 100\%$$

= 91 %

4. KESIMPULAN

Dalam penelitian ini, sentimen publik terhadap polisi Republik Indonesia melalui komentar youtube diklasifikasikan menggunakan metode Support Vector Machine (SVM) dan SMOTE. Data yang digunakan sebanyak 1.000 dataset hasil *crawling*. Berdasarkan hasil eksperimen dan pengujian dengan perbandingan empat split data bahwa rasio 75:25 menghasilkan akurasi tertinggi, yaitu 85,20%, untuk klasifikasi menggunakan SVM tanpa SMOTE. Sedangkan untuk klasifikasi menggunakan SVM berbasis SMOTE pada rasio 70:30 menghasilkan rata-rata akurasi tertinggi, yaitu 91%. Berdasarkan hasil penelitian, dapat disimpulkan bahwa metode SVM berbasis SMOTE dapat diaplikasikan dengan baik dalam analisis sentimen.

DAFTAR RUJUKAN

Amrullah, Ahmad Zuli, Andi Sofyan Anas, Muh Adrian, and Juniarta Hidayat. 2020. "Analisis Sentimen Movie Review Menggunakan Naive Bayes Classifier Dengan Seleksi Fitur Chi Square." *Jurnal* 2(1). https://journal.universitasbumigora.ac.id/index.php/bite.

Analisis Dan Penerapan, . et al. 2019. 7 *Analisis Dan Penerapan Algoritma Support Vector Machine (SVM) Dalam Data Mining Untuk Menunjang Strategi Promosi (Analysis and Application of Algorithm Support Vector Machine (SVM) in Data Mining to Support Promotional Strategies)*.

Athira Luqyana, Wanda, Imam Cholissodin, and Rizal Setya Perdana. 2018. 2 *Analisis Sentimen Cyberbullying Pada Komentar Instagram Dengan Metode Klasifikasi Support Vector Machine*. http://j-ptiik.ub.ac.id.

Deng, Xinyang, Qi Liu, Yong Deng, and Sankaran Mahadevan. 2016. "An Improved Method to Construct Basic Probability Assignment Based on the Confusion Matrix for Classification Problem." *Information Sciences* 340–341: 250–61.

Desai, Mitali, and Mayuri A Mehta. 2016. "Techniques for Sentiment Analysis of Twitter Data: A Comprehensive Survey." In *2016 International Conference on Computing, Communication and Automation (ICCCA)*, , 149–54.

Dwi Fitriani, Reza, Hasbi Yasin, Departemen Statistika, and Fakultas Sains dan Matematika. 2021. "PENANGANAN KLASIFIKASI KELAS DATA TIDAK SEIMBANG DENGAN RANDOM OVERSAMPLING PADA NAIVE BAYES (Studi Kasus: Status Peserta KB IUD Di Kabupaten Kendal)." 10(1): 11–20.

Hidayatullah, Muhammad, Syariful Alam, and Irsan Jaelani. 2 *Sentiment Analysis of Police Performance On Twitter Users Using Naïve Bayes Method*. http://www.kontras.org/index.

Jurnal, Halaman, Harry Andriyan Maulana, and Arita Witanti. 2023. "JURNAL TEKNIK INFORMATIKA AUTOMATIC SUMMARIZING DOKUMEN REPOSITORY DENGAN TERM FREQUENCY-INVERSE DOCUMENT FREQUENCY METHOD." *JUTEKIN* 11(1).

Magnolia, Cindy, Ade Nurhopipah, Dan Bagus, and Adhi Kusuma. 2022. 9 Edu Komputika *Edu Komputika Journal Penanganan Imbalanced Dataset Untuk Klasifikasi Komentar Program Kampus Merdeka Pada Aplikasi Twitter*. http://journal.unnes.ac.id/sju/index.php/edukom.

Munthe, Mampe Parulian, Anton Siswo, Raharjo Ansori, and Reza Rendian Septiawan. *Analisis Sentimen Komentar Pada Saluran Youtube Food Vlogger Berbahasa Indonesia Menggunakan Algoritma Naïve Bayes Sentiment Analysis Comment on Indonesian Youtube Channel About Food Vlogger Using Naïve Bayes Algorithm*.

Rahmawati, Dina, and Rini Silvi. *Pengaruh Faktor Sosial Demografi Dan Kinerja Polisi Terhadap Kepercayaan Masyarakat Kepada Polisi Di Indonesia Tahun 2017 (Influence of Social Demographic Factors and Police Performance on Public Trust in Police in Indonesia in 2017)*.

Sutoyo, Edi et al. "JEPIN (Jurnal Edukasi Dan Penelitian Informatika) Penerapan SMOTE Untuk Mengatasi Imbalance Class Dalam Klasifikasi Television Advertisement Performance Rating Menggunakan Artificial Neural Network."

Syakur, Abdus. 2021. "IMPLEMENTASI METODE LEXICON BASE UNTUK ANALISIS SENTIMEN KEBIJAKAN PEMERINTAH DALAM PENCEGAHAN PENYEBARAN VIRUS CORONA COVID-19 PADA TWITTER." *Jurnal Ilmiah Informatika Komputer* 26(3): 247–60.

Zuber, Konar, and Mahasiswa Program Doktor Ilmu Hukum. 2017. 15 *PERANAN LEMBAGA POLRI DALAM PENEGAKAN HUKUM Oleh*. http://kbbi.web.id/Peranan.