

PELABELAN SEMI-OTOMATIS BERBASIS *INSET LEXICON* DAN *TARGET-ORIENTED* PADA ANALISIS SENTIMEN ISU VASEKTOMI PADA MEDIA SOSIAL X

Nabila Putri Asy Syifa^{1*}, Chaerur Rozikin², Agung Susilo Yuda Irawan³

^{1,2,3}Universitas Singaperbangsa Karawang, Jl.HS.Ronggo Waluyo, Puseurjaya, Telukjambe Timur, Karawang, Jawa Barat 41361, Telp.(0267) 641177

Keywords:

InSet Lexicon, Semi-Automatic Labeling, Sentiment Analysis, Support Vector Machine, Target-Oriented, Vasectomy.

Correspondent Email:

2210631170036@student.unsika.ac.id

Abstrak. Analisis sentimen pada media sosial umumnya memanfaatkan pelabelan otomatis berbasis *lexicon*, namun pendekatan tersebut masih memiliki keterbatasan dalam memahami konteks opini yang mengandung sarkasme, bahasa informal, *code-mixing*, dan istilah domain spesifik. Penelitian ini mengusulkan pendekatan pelabelan semi-otomatis yang menggabungkan Indonesia Sentiment Lexicon (InSet) dan Target-Oriented pada analisis sentimen isu vasektomi di media sosial X (Twitter). Dataset terdiri atas 3.024 opini, dengan 354 data dianotasi secara manual sebagai *ground truth* dan 2.670 data lainnya dilabeli secara otomatis menggunakan InSet. Selanjutnya dilakukan *text preprocessing*, ekstraksi fitur TF-IDF, penyeimbangan kelas menggunakan SMOTE, serta klasifikasi menggunakan Support Vector Machine (SVM) dengan kernel Linear, RBF, Polynomial, dan Sigmoid pada tiga skenario pembagian data. Hasil pengujian menunjukkan bahwa kernel Linear pada skenario 70:30 memberikan performa terbaik dengan Balanced Accuracy 74,09%, Macro Average Precision 73,32%, Macro Average Recall 74,09%, dan Macro Average F1-Score 73,64%. Hasil penelitian menunjukkan bahwa pendekatan pelabelan semi-otomatis mampu menghasilkan dataset berlabel yang lebih efektif untuk mendukung klasifikasi sentimen pada isu vasektomi di media sosial X.



Copyright © 2026 The Author(s), Published by Jurnal Informatika dan Teknik Elektro Terapan. This is an open-access article distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0).

Abstract. Social media sentiment analysis commonly relies on lexicon-based automatic labeling; however, this approach has limitations in understanding opinions containing sarcasm, informal language, code-mixing, and domain-specific terminology. This study proposes a semi-automatic labeling approach that combines the Indonesia Sentiment Lexicon (InSet) and Target-Oriented annotation for sentiment analysis of vasectomy-related discussions on X (Twitter). The dataset consists of 3,024 opinions, of which 354 were manually annotated as the ground truth, while the remaining 2,670 opinions were automatically labeled using the InSet lexicon. The data were subsequently processed through text preprocessing, TF-IDF feature extraction, Synthetic Minority Over-sampling Technique (SMOTE) for class balancing, and classification using Support Vector Machine (SVM) with Linear, RBF, Polynomial, and Sigmoid kernels under three train-test split scenarios. Experimental results show that the Linear kernel with a 70:30 train-test split achieved the best performance, obtaining a Balanced Accuracy of 74.09%, Macro Average Precision of 73.32%, Macro Average Recall of 74.09%, and Macro Average F1-Score of 73.64%. These findings demonstrate that the proposed semi-automatic labeling approach effectively produces a high-quality labeled dataset and supports sentiment classification on vasectomy-related discussions in X (Twitter).

1. PENDAHULUAN

Berdasarkan data Badan Pusat Statistik (BPS), Jawa Barat merupakan provinsi dengan jumlah penduduk terbesar di Indonesia, yaitu mencapai 51.163.888 jiwa pada tahun 2026 [1]. Tingginya jumlah penduduk berdampak pada berbagai permasalahan sosial dan ekonomi, seperti kemiskinan dan stunting [2]. Untuk mengendalikan pertumbuhan penduduk, BKKBN tahun 2025 terus mendorong program Keluarga Berencana (KB). Namun, partisipasi laki-laki masih rendah, yang ditunjukkan oleh jumlah pengguna Metode Operasi Pria (MOP) sebanyak 8.780 orang, jauh di bawah pengguna Metode Operasi Wanita (MOW) sebanyak 89.787 orang di Jawa Barat [3],[4]. Rendahnya partisipasi tersebut dipengaruhi oleh kurangnya pemahaman, stigma sosial, serta anggapan bahwa kontrasepsi merupakan tanggung jawab perempuan.

Kondisi tersebut menjadi perhatian Pemerintah Provinsi Jawa Barat. Pada 29 April 2025, Gubernur Jawa Barat, Kang Dedi Mulyadi (KDM), mengemukakan wacana bahwa peserta program Keluarga Berencana (KB) pria melalui vasektomi dapat diprioritaskan sebagai salah satu syarat penerima bantuan sosial. Menurut KDM, kebijakan tersebut bertujuan untuk meningkatkan keterlibatan laki-laki dalam program KB sekaligus membantu mengendalikan angka kemiskinan rumah tangga [5]. Usulan tersebut memunculkan berbagai tanggapan karena dinilai berkaitan dengan hak reproduksi, kesetaraan gender, dan kebebasan individu dalam menentukan pilihan penggunaan kontrasepsi.

Perdebatan mengenai isu vasektomi kemudian berkembang luas di media sosial X (Twitter) sebagai salah satu platform yang banyak digunakan masyarakat untuk menyampaikan opini. X menjadi ruang diskusi yang memperlihatkan berbagai pandangan, mulai dari dukungan hingga penolakan terhadap usulan tersebut [6], terlebih Indonesia termasuk negara dengan pengguna X terbanyak ketiga di dunia dengan jumlah 25,16 juta pengguna pada tahun 2025 [7]. Tingginya aktivitas ulasan terkait isu vasektomi menghasilkan data tekstual dalam jumlah besar yang dapat dimanfaatkan untuk memahami

kecenderungan opini masyarakat terhadap kebijakan tersebut melalui analisis sentimen, yaitu teknik untuk mengelompokkan opini berdasarkan kecenderungan sentimen tertentu, seperti positif dan negatif.

Analisis sentimen memerlukan algoritma klasifikasi yang andal serta metode pelabelan yang mampu menghasilkan data berkualitas. Penelitian sebelumnya menunjukkan bahwa algoritma Support Vector Machine (SVM) memberikan performa terbaik pada analisis sentimen komentar YouTube mengenai isu Demo DPR RI 2025 dengan akurasi 96,20%, lebih tinggi dibandingkan Neural Network (95,90%) dan Random Forest (88,40%). Namun, pelabelan otomatis berbasis InSet Lexicon masih memiliki keterbatasan dalam menangani *mixed sentiment* dan bahasa tidak baku, sehingga diperlukan pendekatan pelabelan yang lebih mampu memahami konteks opini [8].

Adapun penelitian yang menerapkan Support Vector Machine (SVM) dan K-Nearest Neighbor (KNN) yang dikombinasikan dengan metode SMOTE untuk mengatasi ketidakseimbangan data. Hasilnya yang menunjukkan bahwa kombinasi Support Vector Machine (SVM) dan SMOTE mampu mengatasi ketidakseimbangan kelas dengan menghasilkan Accuracy 81,67%, Precision 85,91%, Recall 79,01%, dan F1-Score 82,32% pada klasifikasi penyakit diabetes. Hasil tersebut menunjukkan bahwa SVM memiliki performa klasifikasi yang baik pada data yang telah diseimbangkan. Penelitian tersebut juga menegaskan pentingnya kualitas data berlabel, di mana pelabelan manual oleh manusia masih memiliki keunggulan karena mampu memahami konteks opini secara lebih utuh dibandingkan pendekatan otomatis berbasis kamus (*lexicon-based*), meskipun membutuhkan waktu dan tenaga yang lebih besar karena proses pelabelan harus dilakukan satu per satu tanpa bantuan sistem komputasi [9].

Berdasarkan tinjauan tersebut, masih terdapat kesenjangan (*gap analysis*) dalam metode pelabelan yang mampu menggabungkan efisiensi *lexicon-based* dengan pemahaman konteks manusia pada opini yang bersifat ambigu, mengandung sarkasme, dan

dipengaruhi bias budaya. Oleh karena itu, kebaruan (*novelty*) penelitian ini mengusulkan pendekatan pelabelan semi-otomatis berbasis InSet Lexicon dan Target-Oriented, melalui pendekatan hibrida ini, proses pelabelan dilakukan dengan mengombinasikan data sampel yang dianotasi secara manual oleh ahli bahasa menggunakan pendekatan berorientasi target (*Target-Oriented*), dengan sisa data lainnya yang dilabeli secara otomatis memanfaatkan kamus *InSet Lexicon*. Dengan skema gabungan ini, aspek ketepatan pemahaman konteks manusia dan kecepatan komputasi *lexicon* dapat berjalan beriringan untuk menghasilkan kualitas label data yang optimal. Dataset hasil pelabelan kemudian diseimbangkan menggunakan Synthetic Minority Over-sampling Technique (SMOTE) dan diklasifikasikan menggunakan Support Vector Machine (SVM) melalui perbandingan empat kernel (Linear, RBF, Polynomial, dan Sigmoid) pada tiga skenario pembagian data (90:10, 80:20, dan 70:30).

2. TINJAUAN PUSTAKA

2.1 Analisis Sentimen

Analisis sentimen merupakan teknik Natural Language Processing (NLP) yang digunakan untuk mengidentifikasi, mengekstraksi, dan mengklasifikasikan opini atau informasi subjektif pada data teks ke dalam kategori sentimen tertentu, seperti positif, negatif, atau netral, sehingga dapat memberikan gambaran mengenai kecenderungan opini yang terkandung dalam suatu dokumen atau kalimat [10].

2.2 Vasektomi

Vasektomi merupakan metode kontrasepsi permanen pada pria yang dilakukan melalui prosedur bedah ringan dengan memutuskan atau mengikat vas deferens untuk mencegah sperma bercampur dengan semen saat ejakulasi sehingga mencegah kehamilan. Prosedur ini tidak memengaruhi produksi hormon testosteron maupun fungsi seksual, sehingga kemampuan ereksi dan ejakulasi tetap normal. Vasektomi juga dikenal sebagai metode kontrasepsi yang aman dan efektif dengan tingkat kegagalan kurang dari 1% [11].

2.3 CRISP-DM

CRISP-DM merupakan metodologi yang menyediakan tahapan terstruktur dalam pelaksanaan proses data mining. Metode ini banyak digunakan karena mampu membantu peneliti mengelola proses analisis data secara sistematis, mulai dari pemahaman masalah hingga evaluasi hasil yang diperoleh [12]. CRISP-DM terdiri atas enam tahapan utama, yaitu Business Understanding untuk memahami kebutuhan dan tujuan bisnis, Data Understanding untuk mengumpulkan serta menganalisis data guna memperoleh wawasan awal dan mengidentifikasi permasalahan kualitas data, Data Preparation untuk menyiapkan data yang siap digunakan melalui proses pembersihan, transformasi, integrasi, dan pemformatan, Modeling untuk membangun dan melatih model *machine learning*, Evaluation untuk mengevaluasi dan memilih model dengan performa terbaik sesuai tujuan penelitian, serta Deployment untuk menerapkan model atau hasil analisis yang telah divalidasi sehingga dapat memberikan manfaat atau nilai pada proses bisnis [13].



Gambar 1. Metode CRISP-DM
(Sumber : (Chapman et al., 2000))

2.4 Pelabelan Inset

Pelabelan Indonesia Sentiment Lexicon (InSet) merupakan pendekatan *lexicon-based* yang menggunakan kamus sentimen bahasa Indonesia yang dikembangkan oleh Fajri Koto dan Gemala Y. Rahmaningtyas. InSet terdiri atas 3.609 kosakata positif dan 6.609 kosakata negatif, dengan nilai polaritas pada rentang -5 hingga $+5$ [14]. Label sentimen ditentukan berdasarkan akumulasi skor polaritas setiap kata, yaitu positif jika skor > 0 , negatif jika skor < 0 , dan netral jika skor $= 0$ [15].

2.5 Pelabelan *Target-Oriented*

Pelabelan *Target-Oriented* merupakan pendekatan anotasi manual yang menentukan label sentimen berdasarkan konteks kalimat dan objek yang menjadi sasaran opini. Pendekatan ini mempertimbangkan unsur seperti negasi, sarkasme, dan hubungan antarkata sehingga mampu menangkap makna implisit secara lebih akurat dibandingkan pendekatan berbasis *lexicon*, serta menghasilkan data berlabel yang lebih berkualitas untuk analisis sentimen [16].

2.6 Text Preprocessing

Text Preprocessing merupakan tahapan awal dalam *text mining* yang bertujuan menghasilkan data teks yang bersih, konsisten, dan siap digunakan pada proses pemodelan [17]. Tahapan yang dilakukan meliputi cleaning untuk menghapus atribut yang tidak diperlukan, seperti emoji, tanda baca, URL, dan *hashtag* [18]. Case folding untuk mengubah seluruh karakter menjadi huruf kecil [19]. Normalization untuk mengubah kata tidak baku atau singkatan menjadi bentuk baku menggunakan kamus *Colloquial Indonesia Lexicon* [20],[21]) selanjutnya normalization foreign word translation untuk menerjemahkan kata atau frasa berbahasa asing, khususnya bahasa Inggris, ke dalam bahasa Indonesia agar konsisten dengan mayoritas kosakata pada dataset [22]. Next word negation untuk mempertahankan informasi negasi dengan menambahkan penanda "NOT_" yang diikuti simbol *underscore* () pada kata setelah negasi, karena berpengaruh terhadap penentuan sentimen, terutama sentimen negatif [23]. Tokenizing untuk memecah teks menjadi token berdasarkan pemisah spasi [24]. Stopword removal untuk menghilangkan kata-kata umum yang kurang memberikan informasi penting dalam proses klasifikasi [25]. Serta stemming untuk mengubah kata berimbuhan menjadi bentuk kata dasar dengan menghilangkan awalan dan akhiran (afiksasi) [26].

2.7 TF-IDF

Term Frequency-Inverse Document Frequency (TF-IDF) merupakan metode ekstraksi fitur yang digunakan untuk memberikan bobot pada setiap kata berdasarkan frekuensi kemunculannya dalam suatu dokumen (*Term Frequency*) dan tingkat kemunculannya pada seluruh kumpulan

dokumen (*Inverse Document Frequency*). Kata yang sering muncul pada suatu dokumen tetapi jarang muncul pada dokumen lain akan memperoleh bobot lebih tinggi sehingga lebih representatif dalam proses klasifikasi [27].

a. Term Frequency (TF)

Term Frequency (TF) digunakan untuk mengukur frekuensi kemunculan suatu istilah (*term*) t di dalam sebuah dokumen d . Nilainya dihitung menggunakan Persamaan (1).

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

Keterangan :

$f_{t,d}$
= Jumlah kemunculan kata t pada dokumen d .

$\sum_k f_{k,d}$
= Total seluruh kata pada dokumen tersebut.

b. Inverse Document Frequency (IDF)

Inverse Document Frequency (IDF) digunakan untuk mengukur tingkat kepentingan suatu istilah berdasarkan frekuensi kemunculannya pada seluruh dokumen dalam korpus. Perhitungannya ditunjukkan pada Persamaan (2).

$$IDF(t) = \log\left(\frac{N}{1+n_t}\right) \quad (2)$$

Keterangan :

N
= Jumlah total dokumen

n_t
= Jumlah dokumen yang mengandung kata t

c. Pembobotan TF-IDF

Inverse Document Frequency (IDF) digunakan untuk mengukur tingkat kepentingan suatu istilah berdasarkan frekuensi kemunculannya pada seluruh dokumen dalam korpus seperti ditunjukkan pada Persamaan (3).

$$TF-IDF(t, d) = TF(t, d) \times IDF(t) \quad (3)$$

Metode TF-IDF telah terbukti mampu mengekstraksi informasi penting dari data teks

secara efektif, sehingga banyak diterapkan pada berbagai tugas *text mining*, seperti klasifikasi dokumen, pencarian informasi (*information retrieval*), dan sistem rekomendasi [27].

2.8 SMOTE

Synthetic Minority Over-sampling Technique (SMOTE) merupakan metode *oversampling* untuk mengatasi ketidakseimbangan kelas (*imbalanced data*) dengan menghasilkan sampel sintetis pada kelas minoritas melalui proses interpolasi antar data yang berdekatan. Pendekatan ini meningkatkan representasi kelas minoritas sehingga mengurangi bias terhadap kelas mayoritas dan membantu meningkatkan performa model klasifikasi [28].

2.9 Pembagian Data (Split Data)

Split data merupakan teknik validasi yang membagi dataset menjadi data latih (*training set*) dan data uji (*testing set*) untuk melatih sekaligus mengevaluasi performa model. Rasio pembagian yang umum digunakan meliputi 70:30, 80:20, dan 90:10, dengan porsi data latih lebih besar agar model mampu mempelajari pola data secara optimal, sedangkan data uji digunakan untuk mengukur kemampuan generalisasi model. Dengan demikian, proses evaluasi dapat dilakukan secara lebih objektif serta membantu meminimalkan risiko overfitting maupun underfitting [29].

2.10 Support Vector Machine (SVM)

Support Vector Machine (SVM) merupakan algoritma *supervised learning* yang banyak digunakan untuk klasifikasi sentimen karena mampu menangani data berdimensi tinggi dengan performa yang stabil. Algoritma ini membangun bidang pemisah (*hyperplane*) dengan *margin* maksimum untuk memisahkan data antar kelas sentimen $y_i \in -1, 1$ pada ruang berdimensi tinggi menggunakan fungsi *kernel*. Oleh karena itu, SVM digunakan sebagai algoritma klasifikasi pada penelitian ini [30]. Berikut persamaannya (4).

$$y_i = W \cdot x_i + b \quad (4)$$

Keterangan:

y_i = nilai kelas dari data ke- i

W = nilai bobot support vector yang tegak lurus dengan

hyperplane b

b = nilai bias atau error

x_i = jumlah atau banyaknya data

Berikut adalah persamaan untuk dua pemisah sejajar (*margin*) dengan hyperplane pada masing-masing kelas yang diperoleh dari support vectors.

Untuk kelas positif:

$$(w \cdot x_i) + b \geq +1, \text{ untuk } y_i = +1 \quad (5)$$

Untuk kelas negatif:

$$(w \cdot x_i) + b \leq -1, \text{ untuk } y_i = -1 \quad (6)$$

Keterangan:

$x_i : [x_1, x_2, x_3]$ adalah vektor baris berdimensi l (banyaknya fitur)

y_i : Label kelas dari x_i

$w : [x_1, x_2, \dots, x_3]$ adalah vektor baris yang berfungsi sebagai parameter bobot

b : Bias atau error

Keunggulan SVM didukung oleh *kernel trick*, yaitu mekanisme pemetaan data ke ruang berdimensi lebih tinggi sehingga pola yang bersifat linear maupun non-linear dapat dipisahkan secara lebih efektif. SVM menyediakan beberapa fungsi *kernel* yang umum digunakan, yaitu Linear, Radial Basis Function (RBF), Polynomial, dan Sigmoid [31]. Berikut penjelasan dari masing-masing kernel:

1) Kernel Linear

Merupakan fungsi *kernel* sederhana pada Support Vector Machine (SVM) yang membentuk fungsi *kernel* berdasarkan hasil *inner product* antara dua vektor fitur, $\langle x, y \rangle$, dengan penambahan konstanta c untuk membentuk fungsi kernel.

2) Kernel Radial Basis Function (RBF)

Merupakan salah satu *kernel* non-linear pada Support Vector Machine (SVM) yang memetakan data ke ruang fitur berdimensi lebih tinggi sehingga mampu memisahkan data yang tidak dapat dipisahkan secara linear. Dibandingkan *kernel* Polynomial, RBF memiliki lebih sedikit *hyperparameter* sehingga lebih sederhana dalam penentuan parameter dan lebih stabil secara komputasi.

3) Kernel Polynomial

Merupakan *kernel* non-linear pada Support Vector Machine (SVM) yang

digunakan untuk memodelkan hubungan data yang kompleks. Kernel ini umumnya memberikan performa yang baik pada data latih yang telah melalui proses normalisasi sehingga distribusi fitur menjadi lebih seragam.

4) Kernel Sigmoid

Merupakan *kernel* non-linear pada Support Vector Machine (SVM) yang dibangun menggunakan fungsi *hyperbolic tangent* (*tanh*). Kernel ini mampu memodelkan hubungan data yang tidak dapat dipisahkan secara linear.

2.11 Evaluation Metrics

Confusion Matrix dan Classification Report merupakan metode evaluasi yang digunakan untuk mengukur kinerja model klasifikasi. *Confusion Matrix* membandingkan hasil prediksi dengan label sebenarnya melalui empat komponen, yaitu True Positive (TP), True Negative (TN), False Positive (FP), dan False Negative (FN). Sementara itu, *Classification Report* menyajikan metrik Accuracy, Precision, Recall, dan F1-Score untuk mengevaluasi performa model secara komprehensif. Pada penelitian ini, evaluasi algoritma Support Vector Machine (SVM) dilakukan menggunakan metrik tersebut [32]. Pada kasus algoritma klasifikasi, Evaluation Metrics yang diukur atau dihitung untuk memahami performa algoritma tersebut adalah sebagai berikut:

1) Accuracy

Merupakan metrik yang mengukur proporsi prediksi benar terhadap seluruh data uji. Semakin tinggi nilai *Accuracy*, semakin baik kinerja model dalam mengklasifikasikan data.

2) Precision

Merupakan metrik yang mengukur ketepatan prediksi kelas positif berdasarkan perbandingan antara prediksi positif yang benar dan seluruh prediksi positif.

3) Recall

Merupakan metrik yang mengukur kemampuan model dalam mengenali seluruh data kelas positif berdasarkan perbandingan antara prediksi positif yang benar dan seluruh data positif.

4) F1-Score

Merupakan rata-rata harmonis dari Precision dan Recall yang memberikan evaluasi seimbang terhadap performa

model, khususnya pada dataset dengan distribusi kelas yang tidak seimbang (*imbalanced data*).

Evaluasi model menggunakan metrik Accuracy, Precision, Recall, dan F1-Score dengan pendekatan Macro Average untuk memberikan bobot yang seimbang pada setiap kelas. Selain itu, validasi dilakukan menggunakan Stratified K-Fold Cross Validation guna menjaga konsistensi distribusi kelas pada setiap *fold* [33].

2.12 Confusion Matrix

Confusion Matrix merupakan metode evaluasi yang membandingkan hasil prediksi model dengan label sebenarnya. Matriks ini disajikan dalam bentuk tabel, di mana baris menunjukkan kelas aktual dan kolom menunjukkan kelas hasil prediksi [34]. Tabel 1 berikut merupakan hasil tabel confusion matrix 2 kelas yaitu positif dan negatif.

Tabel 1. Confusion matrix

Data		Actual Values	
		True	False
Predicted Values	True	TP (True Positive)	FP (False Positive)
	False	FN (False Negative)	TN (True Negative)

Confusion Matrix terdiri atas empat komponen, yaitu True Positive (TP), False Positive (FP), False Negative (FN), dan True Negative (TN). TP menunjukkan data positif yang diprediksi dengan benar, FP menunjukkan data negatif yang diprediksi sebagai positif, FN menunjukkan data positif yang diprediksi sebagai negatif, sedangkan TN menunjukkan data negatif yang diprediksi dengan benar. Keempat komponen tersebut menjadi dasar perhitungan metrik evaluasi, seperti Accuracy, Precision, Recall, dan F1-Score.

2.13 Analisis Thematic

Analisis tematik merupakan metode penelitian kualitatif yang digunakan untuk mengidentifikasi dan mengelompokkan pola atau tema berdasarkan kesamaan makna dalam suatu kumpulan data. Metode ini menghasilkan

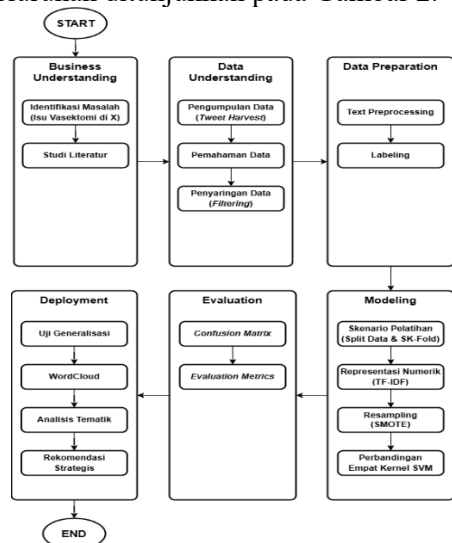
pemahaman yang lebih mendalam terhadap suatu fenomena melalui interpretasi kontekstual, sehingga dapat digunakan sebagai dasar penyusunan rekomendasi strategis [35].

2.14 WorldCloud

WordCloud merupakan visualisasi kumpulan teks yang menampilkan frekuensi kemunculan suatu kata melalui perbedaan ukuran dan posisinya. Kata-kata yang memiliki frekuensi kemunculan lebih tinggi akan ditampilkan dengan ukuran yang lebih besar dibandingkan kata lainnya [36].

3. METODE PENELITIAN

Penelitian ini menerapkan metodologi Cross Industry Standard Process for Data Mining (CRISP-DM) yang terdiri atas enam tahapan, yaitu *Business Understanding*, *Data Understanding*, *Data Preparation*, *Modeling*, *Evaluation*, dan *Deployment*. Metodologi tersebut dipilih karena menyediakan alur kerja yang sistematis dalam proses pengolahan data hingga pembangunan model klasifikasi sentimen. Fokus penelitian ini berada pada penerapan pelabelan semi-otomatis yang mengombinasikan pelabelan manual berbasis Target-Oriented pada data sampel dengan pelabelan otomatis menggunakan Indonesia Sentiment Lexicon (InSet) pada sisa populasi data. Dataset hasil pelabelan kemudian digunakan pada proses pemodelan menggunakan algoritma Support Vector Machine (SVM) untuk mengevaluasi performa klasifikasi sentimen. Alur penelitian secara keseluruhan ditunjukkan pada Gambar 2.



Gambar 2. Alur penelitian

3.1 Business Understanding

Tahap *Business Understanding* dilakukan untuk mengidentifikasi permasalahan penelitian berdasarkan berkembangnya opini masyarakat di media sosial X mengenai isu vasektomi di Jawa Barat setelah pernyataan Gubernur Jawa Barat, Dedi Mulyadi. Beragam tanggapan yang muncul menunjukkan adanya perbedaan pandangan masyarakat terhadap isu tersebut sehingga diperlukan analisis sentimen untuk mengetahui kecenderungan opini publik. Oleh karena itu, penelitian ini menerapkan pendekatan pelabelan semi-otomatis berbasis Target-Oriented dan Indonesia Sentiment Lexicon (InSet) serta algoritma Support Vector Machine (SVM) guna memperoleh model klasifikasi sentimen dengan performa yang optimal.

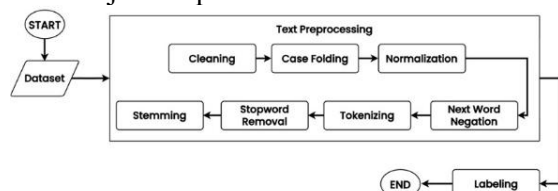
3.2 Data Understanding

Dataset penelitian diperoleh dari platform media sosial X (Twitter) menggunakan perangkat lunak Tweet Harvest (v2.6.1) yang dijalankan pada lingkungan Google Colaboratory. Pengambilan data dilakukan berdasarkan kata kunci "vasektomi", "jawa barat", dan "jabar" dalam rentang waktu 29 April 2025 hingga 31 Juli 2025. Periode tersebut dipilih karena 29 April 2025 merupakan waktu ketika pernyataan Gubernur Jawa Barat, Dedi Mulyadi, mengenai usulan vasektomi sebagai salah satu syarat penerima bantuan sosial mulai menjadi perhatian publik dan memicu berbagai diskusi di media sosial X.

Selain itu, pengambilan data juga mempertimbangkan berbagai unggahan dan tanggapan masyarakat yang membahas isu tersebut sehingga mampu merepresentasikan perkembangan opini publik selama periode pengamatan. Selanjutnya, dataset dipahami karakteristiknya melalui identifikasi jumlah data, tipe data, atribut yang tersedia, serta kualitas data dengan melakukan pemeriksaan terhadap nilai *null*, data duplikat, dan data yang tidak relevan. Tahapan ini bertujuan untuk memperoleh pemahaman mengenai kondisi dataset sebagai dasar pelaksanaan *text preprocessing* dan proses analisis pada tahap berikutnya.

3.3 Data Preparation

Tahap *Data Preparation* bertujuan menyiapkan dataset agar siap digunakan pada proses pemodelan. Pada tahap ini dilakukan serangkaian proses *text preprocessing* untuk membersihkan, menyeragamkan, dan mengurangi *noise* pada data teks sehingga kualitas data menjadi lebih baik. Selanjutnya, dilakukan proses pelabelan semi-otomatis dengan mengombinasikan pelabelan manual berbasis *Target-Oriented* pada data sampel dan pelabelan otomatis menggunakan *Indonesia Sentiment Lexicon* (InSet) pada sisa populasi data. Hasil dari kedua proses tersebut kemudian digabungkan menjadi satu dataset berlabel yang digunakan sebagai masukan pada tahap pemodelan. Tahapan *Data Preparation* yang diterapkan pada penelitian ini ditunjukkan pada Gambar 3.



Gambar 3. Alur Preparation

a) Cleaning Data

Tahap *cleaning* dilakukan untuk menghilangkan komponen yang tidak diperlukan pada data teks sehingga diperoleh data yang lebih bersih sebelum diproses lebih lanjut. Proses ini meliputi penghapusan URL, username (@mention), hashtag (#), emoji, angka, tanda baca, karakter khusus, serta spasi berlebih yang tidak memberikan informasi terhadap analisis sentimen.

b) Case Folding

Tahap *case folding* dilakukan untuk menyeragamkan seluruh karakter pada data teks menjadi huruf kecil (*lowercase*) sehingga tidak terjadi perbedaan representasi kata akibat penggunaan huruf kapital.

c) Normalization

Tahap *normalization* dilakukan untuk mengubah kata tidak baku, singkatan, dan bahasa gaul menjadi bentuk baku agar representasi kata menjadi lebih konsisten. Proses normalisasi menerapkan pendekatan kombinasi (*hybrid normalization*) menggunakan Colloquial Indonesian

Lexicon yang dipadukan dengan kamus kustom (*custom dictionary*) yang disusun berdasarkan pedoman Kamus Besar Bahasa Indonesia (KBBI) serta disesuaikan dengan kosakata yang sering muncul pada pembahasan isu vasektomi di media sosial X. Selain itu, penelitian ini juga menerapkan normalisasi *foreign word translation* menggunakan Google Translate API untuk menerjemahkan kata atau frasa berbahasa asing ke dalam bahasa Indonesia sehingga diperoleh representasi teks yang lebih beragam sebelum memasuki tahap berikutnya.

d) Next Word Negation

Tahap *next word negation* dilakukan untuk mempertahankan informasi negasi yang terdapat pada teks agar makna sentimen tidak berubah selama proses analisis. Proses ini dilakukan dengan menambahkan awalan "NOT_" pada kata yang berada setelah kata negasi, seperti *tidak*, *bukan*, *jangan*, dan *belum*, sehingga hubungan antara kata negasi dan kata yang mengikutinya tetap dipertahankan. Tahapan ini diterapkan untuk membantu model membedakan kata yang memiliki makna positif maupun negatif ketika dipengaruhi oleh konteks negasi.

e) Tokenizing

Tahap *tokenizing* dilakukan untuk memecah setiap kalimat menjadi kumpulan token atau kata berdasarkan karakter spasi sebagai pemisah. Proses ini bertujuan mengubah data teks menjadi satuan kata yang lebih terstruktur sehingga memudahkan proses identifikasi kosakata

f) Stopword Removal

Tahap *stopword removal* dilakukan untuk menghilangkan kata-kata yang memiliki frekuensi kemunculan tinggi namun memberikan kontribusi yang rendah terhadap proses klasifikasi sentimen, seperti kata hubung, kata depan, dan kata ganti. Proses ini menggunakan daftar *stopword* bahasa Indonesia yang telah disesuaikan dengan karakteristik data penelitian, sehingga kata-kata yang masih memiliki informasi penting terhadap analisis sentimen tetap dipertahankan. Selain itu, kata-kata negasi yang telah diproses pada tahap *next word negation* tidak dihapus agar informasi negasi tetap

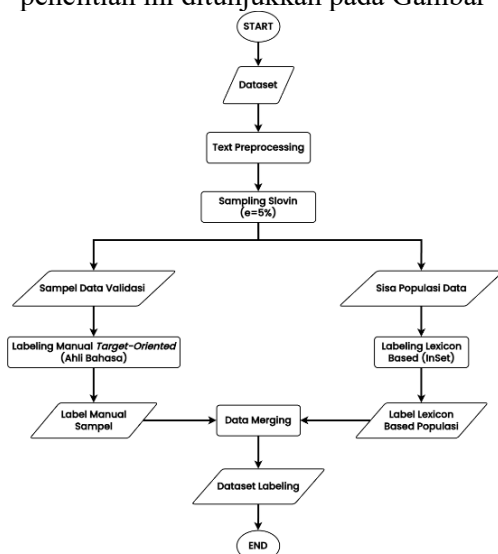
terjaga dan makna sentimen pada setiap opini tidak berubah.

g) *Stemming*

Tahap *stemming* dilakukan untuk mengubah setiap kata berimbuhan menjadi bentuk kata dasar sehingga variasi penulisan kata yang memiliki makna sama dapat direpresentasikan secara konsisten. Proses ini menerapkan algoritma *stemming* bahasa Indonesia menggunakan pustaka Sastrawi untuk menghilangkan awalan, akhiran, sisipan, maupun konfiks sesuai kaidah morfologi bahasa Indonesia. Hasil dari tahap *stemming* berupa kumpulan kata dasar yang selanjutnya digunakan pada proses pelabelan sentimen.

h) *Labelling*

Tahap *labeling* dilakukan menggunakan pendekatan semi-otomatis yang mengombinasikan pelabelan manual berbasis Target-Oriented dan pelabelan otomatis menggunakan Indonesia Sentiment Lexicon (InSet). Alur pelabelan semi-otomatis yang diterapkan pada penelitian ini ditunjukkan pada Gambar 4.



Gambar 4. Alur Labelling

Proses pelabelan diawali dengan penentuan sampel menggunakan rumus Slovin dengan tingkat kesalahan (*error tolerance*) sebesar 5%. Sampel yang diperoleh selanjutnya diberi label secara manual oleh ahli bahasa menggunakan pendekatan Target-Oriented, yaitu pelabelan yang mempertimbangkan konteks kalimat, sasaran pembahasan (*target*), serta makna keseluruhan opini terhadap isu

yang dibahas [16]. Setelah proses tersebut, sisa populasi data diberi label secara otomatis menggunakan InSet Lexicon berdasarkan skor polaritas setiap kata pada teks. Selanjutnya, hasil pelabelan manual dan pelabelan otomatis digabungkan melalui proses *data merging* sehingga menghasilkan satu dataset berlabel yang siap digunakan pada tahap pemodelan.

3.4 Modeling

3.4.1 *Transformasi Data*

Setelah tahap *Data Preparation* selesai, data teks ditransformasikan ke dalam bentuk representasi numerik menggunakan metode Term Frequency–Inverse Document Frequency (TF-IDF) sebelum digunakan pada proses pelatihan model. Pembobotan TF-IDF diterapkan untuk menghasilkan vektor fitur berdasarkan tingkat kepentingan setiap kata terhadap keseluruhan korpus.

3.4.2 *SMOTE*

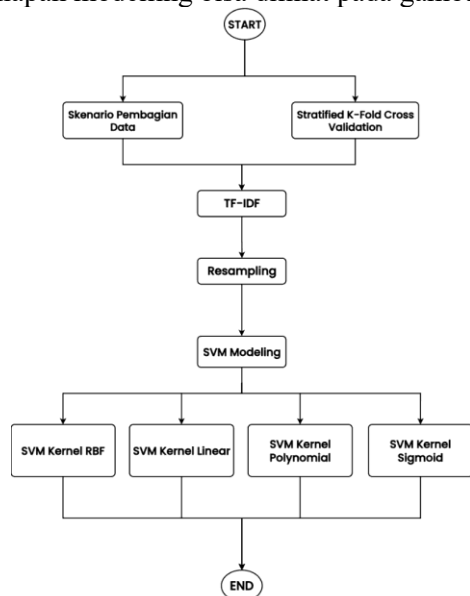
Setelah data direpresentasikan menggunakan TF-IDF, dilakukan penyeimbangan distribusi kelas menggunakan Synthetic Minority Over-sampling Technique (SMOTE) pada data latih. Teknik ini diterapkan untuk mengatasi ketidakseimbangan jumlah data antara kelas sentimen positif dan negatif sehingga proses pelatihan model menjadi lebih optimal. Penerapan SMOTE dilakukan pada kedua skenario pemodelan, yaitu *data split* dan Stratified K-Fold Cross Validation, dengan tetap menerapkan resampling hanya pada data latih untuk mencegah terjadinya *data leakage*.

3.4.3 *Support Vector Machine (SVM)*

Tahap *Modeling* dilakukan menggunakan algoritma Support Vector Machine (SVM) untuk mengklasifikasikan sentimen opini masyarakat terhadap isu vasektomi pada media sosial X ke dalam kelas sentimen positif dan negatif. Proses pemodelan terdiri atas dua skenario pengujian, yaitu pembagian data (*data split*) dan Stratified K-Fold Cross Validation. Pada skenario *data split*, digunakan tiga variasi rasio data latih dan data uji, yaitu 90:10, 80:20, dan 70:30, sedangkan pada skenario kedua digunakan Stratified K-Fold Cross Validation dengan 10 fold.

Pada kedua skenario tersebut, proses pelatihan dilakukan menggunakan empat jenis kernel Support Vector Machine, yaitu Linear,

Radial Basis Function (RBF), Polynomial, dan Sigmoid dengan parameter bawaan (*default*), khususnya pada parameter C dan γ (gamma). Konfigurasi tersebut digunakan untuk membandingkan performa masing-masing kernel secara objektif sehingga dapat diperoleh model dengan tingkat akurasi, stabilitas, dan kemampuan generalisasi yang paling optimal. Tahapan modelling bisa dilihat pada gambar 4.



Gambar 4. Alur Modelling

3.5 Evaluation

Tahap *Evaluation* dilakukan untuk mengukur performa model Support Vector Machine (SVM) dalam mengklasifikasikan sentimen opini masyarakat terkait isu vasektomi pada media sosial X. Evaluasi dilakukan terhadap seluruh hasil eksperimen pada skenario *data split* maupun Stratified K-Fold Cross Validation menggunakan Confusion Matrix sebagai dasar perhitungan kinerja model.

Berdasarkan nilai True Positive (TP), True Negative (TN), False Positive (FP), dan False Negative (FN), performa model diukur menggunakan metrik Balanced Accuracy, Precision, Recall, dan F1-Score. Mengingat dataset memiliki distribusi kelas yang tidak seimbang, perhitungan Precision, Recall, dan F1-Score menggunakan pendekatan Macro Average agar setiap kelas memperoleh bobot evaluasi yang setara, sedangkan Balanced Accuracy digunakan untuk memberikan ukuran performa yang lebih objektif pada kondisi *imbalanced data* [37]. Hasil evaluasi

tersebut selanjutnya digunakan untuk menentukan kombinasi kernel dan skenario pelatihan yang menghasilkan performa klasifikasi terbaik.

3.6 Deployment

Tahap *Deployment* dilakukan setelah proses pemodelan dan evaluasi selesai untuk menguji kemampuan generalisasi algoritma Support Vector Machine (SVM) dengan kernel terbaik terhadap data baru (*unseen data*). Pengujian dilakukan menggunakan kumpulan tweet terbaru mengenai isu vasektomi di media sosial X yang berada di luar rentang waktu dataset utama sehingga dapat menggambarkan kemampuan model dalam mengklasifikasikan opini pada kondisi nyata.

Hasil klasifikasi selanjutnya divisualisasikan menggunakan WordCloud berdasarkan kategori sentimen positif dan negatif untuk mengidentifikasi kata-kata yang paling dominan muncul. Temuan tersebut kemudian dianalisis menggunakan pendekatan analisis tematik guna mengidentifikasi pola opini masyarakat dan disintesis menjadi rekomendasi strategis berbasis data sebagai masukan bagi instansi terkait dalam menyusun strategi edukasi dan penyebaran informasi mengenai program vasektomi.

4. HASIL DAN PEMBAHASAN

4.1 Business Understanding

Tahap *Business Understanding* dilakukan untuk menentukan fokus penelitian, yaitu analisis sentimen opini masyarakat pada media sosial X mengenai isu vasektomi di Provinsi Jawa Barat. Berdasarkan tujuan tersebut, penelitian mengumpulkan 3.127 tweet sebagai dataset yang diproses menggunakan pendekatan pelabelan semi-otomatis dengan mengombinasikan pelabelan manual berbasis *Target-Oriented* pada data sampel dan pelabelan otomatis berbasis InSet Lexicon pada sisa populasi data untuk menghasilkan dataset berlabel sentimen positif dan negatif. Dataset tersebut selanjutnya digunakan dalam proses pemodelan menggunakan algoritma *Support Vector Machine* (SVM) melalui tiga skenario pembagian data (*data split*), empat jenis kernel SVM (*Linear*, *RBF*, *Polynomial*, dan *Sigmoid*), serta validasi *Stratified K-Fold Cross Validation* guna memperoleh model dengan performa dan kemampuan generalisasi terbaik.

Hasil klasifikasi kemudian dimanfaatkan sebagai dasar analisis tematik dalam penyusunan rekomendasi strategis terkait isu vasektomi.

4.2 Data Understanding

Tahap *Data Understanding* dilakukan untuk memperoleh dan memahami karakteristik dataset yang digunakan dalam penelitian. Data dikumpulkan melalui proses *scraping* pada media sosial X menggunakan Tweet Harvest (v2.6.1) di lingkungan Google Colaboratory dengan kata kunci "*vasektomi*", "*Jawa Barat*", dan "*Jabar*" pada periode 29 April 2025 hingga 31 Juli 2025. Hasil *scraping* memperoleh 3.127 tweet yang merepresentasikan opini masyarakat mengenai isu vasektomi di Provinsi Jawa Barat.

Dataset yang diperoleh terdiri atas 15 atribut, namun penelitian ini hanya menggunakan atribut *full_text* sebagai data utama untuk analisis sentimen. Selanjutnya dilakukan pemeriksaan karakteristik dataset, meliputi jumlah data, tipe data, nilai *null*, dan data duplikat sebagai dasar proses *Data Preparation*.

4.3 Data Preparation

Tahap *Data Preparation* bertujuan menghasilkan dataset yang bersih dan siap digunakan pada proses klasifikasi sentimen. Pada tahap ini dilakukan serangkaian *text preprocessing* untuk membersihkan, menyeragamkan, dan mengurangi *noise* pada data teks, yang meliputi *Cleaning*, *Case Folding*, *Normalization (hybrid normalization dan foreign word translation)*, *Next Word Negation*, *Tokenizing*, *Stopword Removal*, dan *Stemming*. Selanjutnya, diterapkan pelabelan semi-otomatis dengan mengombinasikan pelabelan manual berbasis *Target-Oriented* pada data sampel dan pelabelan otomatis menggunakan *Indonesia Sentiment Lexicon (InSet)* pada sisa populasi data. Tahapan yang dilakukan adalah sebagai berikut:

1. Cleaning Data

Cleaning dilakukan untuk menghapus data yang tidak valid serta berbagai elemen *noise* seperti *mention*, *hashtag*, URL, angka, tanda baca, emoji, karakter berulang, dan spasi berlebih. Selain itu, data duplikat dan data kosong hasil proses *cleaning* juga dieliminasi sehingga diperoleh 3.024 data

yang siap digunakan pada tahapan berikutnya.

2. Case Folding

Case Folding dilakukan dengan mengubah seluruh huruf kapital menjadi huruf kecil (*lowercase*) untuk menyeragamkan representasi kata pada dataset. Proses ini memastikan kata yang memiliki makna sama tetapi berbeda penulisan huruf diperlakukan sebagai satu token yang identik.

3. Normalization

Normalization dilakukan untuk menyeragamkan penulisan kata agar memiliki bentuk yang konsisten sebelum proses ekstraksi fitur. Pada penelitian ini digunakan pendekatan *hybrid normalization* yang mengombinasikan *Colloquial Indonesian Lexicon*, *custom dictionary* yang disusun sesuai domain penelitian, serta *Foreign Word Translation* menggunakan *Google Translate API*. Pendekatan ini digunakan untuk memperbaiki kata tidak baku, singkatan, istilah domain, dan kata berbahasa asing sehingga menghasilkan representasi teks yang lebih seragam.

4. Next Word Negation

Next Word Negation dilakukan untuk mempertahankan makna negasi pada opini sebelum memasuki tahap *Stopword Removal*. Tahap ini mengidentifikasi kata negasi seperti *tidak*, *bukan*, *jangan*, *belum*, *enggak*, *ga*, *gak*, *tak*, dan *tanpa*, kemudian menggabungkannya dengan kata yang mengikutinya menggunakan penanda *NOT_*. Pendekatan ini diterapkan untuk mencegah hilangnya kata negasi pada proses *Stopword Removal* sekaligus mempertahankan konteks sentimen saat pelabelan otomatis menggunakan *InSet Lexicon*, sehingga makna opini tetap terjaga.

5. Tokenizing

Tokenizing dilakukan untuk memecah setiap kalimat menjadi kumpulan kata tunggal (*token*) berdasarkan spasi (*whitespace*). Proses ini bertujuan mengubah data teks menjadi unit kata yang lebih mudah diproses pada tahapan *Stopword Removal*, *Stemming*, serta ekstraksi fitur menggunakan *TF-IDF*.

6. Stopword Removal

Stopword Removal dilakukan untuk menghapus kata-kata umum (*stopwords*)

yang tidak memberikan kontribusi signifikan terhadap proses klasifikasi, seperti kata hubung, kata depan, dan kata ganti. Proses ini menggunakan daftar *stopwords* dari library Sastrawi. Berbeda dengan *stopwords* lainnya, token hasil Next Word Negation yang memiliki awalan NOT_ tetap dipertahankan sehingga makna negasi pada opini tidak hilang selama proses *preprocessing*.

7. Stemming

Stemming merupakan tahap akhir pada proses *text preprocessing* yang bertujuan mengubah setiap token menjadi bentuk kata dasar dengan menghilangkan imbuhan. Proses ini dilakukan untuk menyeragamkan variasi kata yang memiliki akar kata sama sehingga mengurangi redundansi fitur dan meningkatkan konsistensi representasi teks pada proses klasifikasi.

Hasil dari tahapan diatas dapat dilihat pada tabel 2.

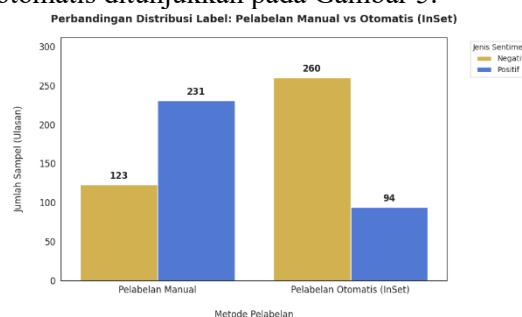
Tabel 2. Hasil Preparation

Proses	Hasil
Data mentah	Informasi HOAX vasektomi tdk membuat loyo ejakulasi yg dipotong hny saluran sperma saja Ereksi dan vitalitas tetap sama
Case folding	informasi hoax vasektomi tdk membuat loyo ejakulasi yg dipotong hny saluran sperma saja ereksi dan vitalitas tetap sama
Normalisasi hybrid	informasi hoax vasektomi tidak membuat loyo ejakulasi yang dipotong hanya saluran sperma saja ereksi dan vitalitas tetap sama
Normalisasi foreign word translation	informasi lelucon vasektomi tidak membuat loyo ejakulasi yang dipotong hanya saluran sperma saja ereksi dan vitalitas tetap sama
Next word negation	informasi lelucon vasektomi NOT_ membuat loyo ejakulasi yang dipotong hanya saluran sperma saja ereksi dan vitalitas tetap sama
Tokenizing	['informasi', 'lelucon', 'vasektomi',

	['NOT_ membuat', 'loyo', 'ejakulasi', 'yang', 'dipotong', 'hanya', 'saluran', 'sperma', 'saja', 'ereksi', 'dan', 'vitalitas', 'tetap', 'sama']
Stopword removal	['informasi', 'lelucon', 'vasektomi', 'NOT_ membuat', 'loyo', 'ejakulasi', 'dipotong', 'saluran', 'sperma', 'ereksi', 'vitalitas', 'tetap', 'sama']
Stemming	['informasi', 'lelucon', 'vasektomi', 'NOT_ buat', 'loyo', 'ejakulasi', 'potong', 'saluran', 'sperma', 'ereksi', 'vitalitas', 'tetap', 'sama']

4.4 Labelling

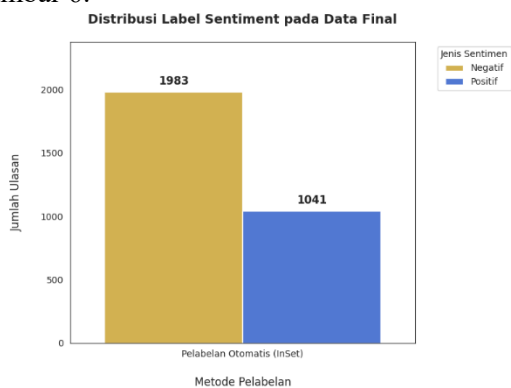
Tahap *labeling* menggunakan pendekatan pelabelan semi-otomatis dengan mengombinasikan pelabelan manual berbasis Target-Oriented dan pelabelan otomatis menggunakan Indonesia Sentiment Lexicon (InSet) [14],[16] sebanyak 354 data sampel dianotasi secara manual sebagai *ground truth*, kemudian divalidasi menggunakan InSet Lexicon. Hasil validasi menunjukkan akurasi sebesar 39,83%, yang mengindikasikan bahwa pendekatan *lexicon-based* masih memiliki keterbatasan dalam memahami konteks opini, seperti sarkasme, istilah medis, bahasa informal, dan *code-mixing*. Temuan ini sejalan dengan [9] yang menyatakan bahwa pelabelan manual lebih mampu memahami konteks dibandingkan pelabelan berbasis kamus. Perbandingan hasil pelabelan manual dan otomatis ditunjukkan pada Gambar 5.



Gambar 5. Perbandingan Distribusi Kelas Hasil Pelabelan Manual vs Otomatis

Selanjutnya, pelabelan otomatis diterapkan pada 2.670 data populasi, kemudian digabungkan dengan 354 data hasil pelabelan manual sehingga diperoleh 3.024 data berlabel.

Pendekatan ini mengimplementasikan pelabelan semi-otomatis berbasis InSet Lexicon dan Target-Oriented sebagai kebaruan penelitian, sehingga efisiensi pelabelan tetap terjaga tanpa mengurangi kualitas label. Distribusi kelas dataset akhir ditunjukkan pada Gambar 6.



Gambar 6. Distribusi Kelas Dataset Hasil Pelabelan Semi-Otomatis

Dataset hasil pelabelan masih menunjukkan kondisi *class imbalance*, sehingga pada tahap berikutnya diterapkan Synthetic Minority Over-sampling Technique (SMOTE) sebelum proses pemodelan menggunakan algoritma Support Vector Machine (SVM).

4.5 Modelling

4.5.1 Splitting Data

Setelah proses pelabelan selesai dilakukan, dataset hasil pelabelan sebanyak 3.024 data dibagi menjadi data latih (*training data*) dan data uji (*testing data*). Penelitian ini menggunakan tiga skenario pembagian data, yaitu 90:10, 80:20, dan 70:30, untuk menganalisis pengaruh proporsi data latih terhadap performa algoritma Support Vector Machine (SVM). Rincian pembagian data pada setiap skenario ditunjukkan pada Tabel 3.

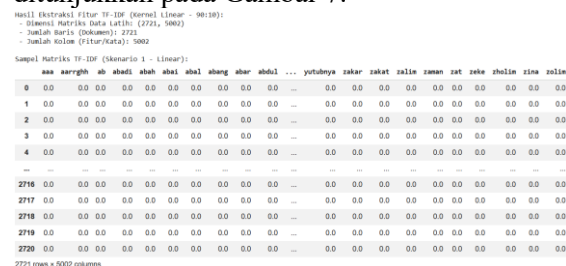
Tabel 3. Rincian Pembagian Data Latih dan Data Uji

Skenario	Data Latih	Data Uji	Total Data
90:10	2.721	303	3.024
80:20	2.419	605	3.024
70:30	2.116	908	3.024

4.5.2 TF-IDF

Data latih dan data uji pada setiap skenario pembagian data selanjutnya ditransformasikan ke dalam bentuk vektor numerik menggunakan metode Term Frequency-Inverse Document

Frequency (TF-IDF) melalui modul TfidfVectorizer pada *library* scikit-learn. Untuk menghindari *data leakage*, proses *fit-transform* hanya diterapkan pada data latih, sedangkan data uji hanya melalui proses *transform*. Hasil pembobotan TF-IDF menghasilkan sparse matrix sebagai representasi numerik setiap dokumen berdasarkan bobot kata. Jumlah fitur yang dihasilkan berbeda pada setiap skenario, yaitu 5.002 fitur pada skenario 90:10, 4.695 fitur pada skenario 80:20, dan 4.382 fitur pada skenario 70:30. Hasil transformasi TF-IDF ditunjukkan pada Gambar 7.



Gambar 7. Hasil TF-IDF

Representasi numerik yang dihasilkan pada tahap ini selanjutnya digunakan sebagai masukan pada proses Synthetic Minority Over-sampling Technique (SMOTE) untuk menyeimbangkan distribusi kelas sebelum dilakukan pelatihan model Support Vector Machine (SVM).

4.5.3 SMOTE

Distribusi kelas pada dataset hasil pelabelan menunjukkan kondisi *class imbalance*, sehingga diterapkan Synthetic Minority Over-sampling Technique (SMOTE) menggunakan *library* imbalanced-learn (*random_state=42*). Proses *resampling* dilakukan hanya pada data latih hasil transformasi TF-IDF pada setiap skenario pembagian data. Hasil penerapan SMOTE ditunjukkan pada Tabel 4.

Tabel 4. Distribusi Data Latih Sebelum dan Sesudah SMOTE

Skenario	Sebelum SMOTE (Positif)	Sebelum SMOTE (Negatif)	Sesudah SMOTE (Positif)	Sesudah SMOTE (Negatif)
90:10	937	1.784	1.784	1.784
80:20	833	1.586	1.586	1.586
70:30	728	1.388	1.388	1.388

Berdasarkan Tabel 4, SMOTE berhasil menyeimbangkan distribusi kelas pada seluruh skenario sehingga dataset siap digunakan pada proses pelatihan model Support Vector Machine (SVM).

4.6 Evaluation

Performa model dievaluasi menggunakan Balanced Accuracy, Macro Average Precision, Macro Average Recall, dan Macro Average F1-Score pada tiga skenario pembagian data. Hasil evaluasi setiap kernel ditunjukkan pada Tabel 5.

Tabel 5. Hasil Evaluation Metrics Seluruh Skenario

Jenis Kernel SVM (TF-IDF)	Skenario	Balanced Accuracy	Macro Avg Precision	Macro Avg Recall	Macro Avg F1-Score
Linear	Skenario 2 (80:20)	0.7323	0.7293	0.7323	0.7307
	Skenario 3 (70:30)	0.7409	0.7332	0.7409	0.7364
	Skenario 1 (90:10)	0.7191	0.7153	0.7191	0.7171
RBF	Skenario 1 (90:10)	0.6761	0.7470	0.6761	0.6877
	Skenario 2 (80:20)	0.6605	0.7367	0.6605	0.6699
	Skenario 3 (70:30)	0.6627	0.7426	0.6627	0.6724
Polynomial	Skenario 2 (80:20)	0.5937	0.7042	0.5937	0.5831
	Skenario 1 (90:10)	0.5752	0.6479	0.5752	0.5628
	Skenario 3 (70:30)	0.5872	0.6880	0.5872	0.5748
Sigmoid	Skenario 2 (80:20)	0.7285	0.7227	0.7285	0.7252
	Skenario 3 (70:30)	0.7390	0.7284	0.7390	0.7324
	Skenario 1 (90:10)	0.7045	0.7009	0.7045	0.7025

Berdasarkan Tabel 5, kernel Linear menghasilkan performa terbaik dengan Macro Average F1-Score tertinggi sebesar 0,7364 pada skenario 70:30, diikuti Sigmoid, RBF, dan Polynomial. Meskipun RBF memperoleh Macro Average Precision tertinggi, nilai Macro Average F1-Score masih berada di bawah kernel Linear dan Sigmoid. Untuk menguji konsistensi performa model, dilakukan Stratified K-Fold Cross Validation (K=10). Hasil pengujian ditunjukkan pada Tabel 6.

Tabel 6. Hasil Stratified K-Fold Cross Validation

Jenis Kernel SVM (TF-IDF)	Balanced Accuracy (Mean)	Macro Avg Precision (Mean)	Macro Avg Recall (Mean)	Macro Avg F1-Score (Mean)
Linear	0.7305	0.7270	0.7305	0.7278
Sigmoid	0.7260	0.7187	0.7260	0.7205
RBF	0.6679	0.7531	0.6679	0.6780
Polynomial	0.5867	0.6945	0.5867	0.5732

Hasil validasi menunjukkan pola yang konsisten, di mana kernel Linear kembali memperoleh performa terbaik dengan Macro Average F1-Score sebesar 0,7278, diikuti Sigmoid (0,7205), RBF (0,6780), dan Polynomial (0,5732). Hasil ini menunjukkan bahwa kernel Linear merupakan model yang paling optimal dan stabil untuk klasifikasi sentimen isu vasektomi.

4.7 Deployment

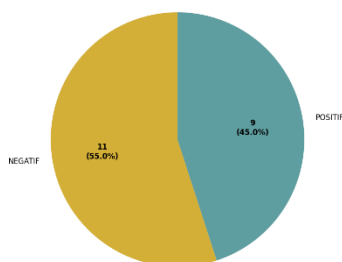
Tahap *deployment* menggunakan model terbaik, yaitu Support Vector Machine (SVM) dengan kernel Linear pada skenario 70:30. Model diuji menggunakan 20 opini terbaru mengenai isu vasektomi di media sosial X yang tidak terlibat pada proses pelatihan maupun pengujian (*unseen data*) untuk mengevaluasi kemampuan generalisasi model. Hasil prediksi sentimen ditunjukkan pada Tabel 7.

Tabel 7. Hasil Prediksi Sentimen Opini Terbaru

No	Opini Asli (Raw Tweet)	Hasil Prediksi Sentimen
1	@jabarnews Menurut gue masalah kemiskinan di Jawa Barat itu bukan km bapak-bapak belum vasektomi. Pemerintah harus fokus buka lapangan kerja dulu daripada terus dorong program vasektomi.	NEGATIF
2	@kangdedi_id Gue SETUJU edukasi kesehatan reproduksi buat pria harus diperluas. Banyak warga yg masih salah paham soal vasektomi padahal tujuannya buat kesejahteraan keluarga.	NEGATIF
3	Jujur aja daripada ribut soal halal haram di medsos, mending fokus kasih edukasi kesehatan reproduksi dan pengendalian penduduk yg benar ke masyarakat Jawa Barat.	NEGATIF
4	@wargajabar KDM boleh aja kampanye vasektomi, tp menurut gue itu BUKAN solusi utama buat ningkatin kesejahteraan masyarakat Jawa Barat.	NEGATIF
5	Budaya banyak anak banyak rezeki udah ga relevan lagi. Tapi menurut gue pemerintah juga jangan cuma jual program vasektomi doang.	POSITIF
6	@sumisuc Edukasi keluarga berencana buat pria emang penting, tp pendekatan pemerintah soal vasektomi masih kurang meyakinkan buat masyarakat desa.	POSITIF
7	Daripada sibuk debat vasektomi, mending pemerintah fokus ngurain kemiskinan dan buka lapangan kerja buat warga Jawa Barat.	NEGATIF
8	Edukasi soal vasektomi penting sih, tp gue pribadi terduga SETUJU kalo prosedur ini diadain solusi utama masalah kependudukan.	NEGATIF
9	Pengangguran makin tinggi, ekonomi makin susah, eh yg dibahas malah vasektomi terus wkwk. Padahal masalah utamanya bukan di situ.	NEGATIF
10	@masyarifto Lelakan penduduk memang perlu dikendalikan, tp kebijakan vasektomi masih menimbulkan banyak pro kontra di masyarakat Jawa Barat.	POSITIF
11	Menurut gue sosialisasi yg baik bisa ngurangin stigma masyarakat terhadap vasektomi. Banyak orang sebenarnya belum paham manfaat kesehatannya.	POSITIF
12	Faktor budaya dan pandangan konservatif warga Jawa Barat bikin program vasektomi susah diterima. Realitanya emang gitu.	POSITIF
13	Masih banyak warga yg nganggap vasektomi itu serem bahkan disainin sama kebiri. Pemerintah harus beresin miskonsepsi ini dulu.	POSITIF
14	Kondisi ekonomi lagi susah, harga kebutuhan naik, lapangan kerja terbatas. Masyarakat lebih butuh kesejahteraan daripada kampanye vasektomi terus-terusan.	NEGATIF
15	@politikjabar Gue paham tujuan vasektomi buat ngurangin kemiskinan, tp menurut gue solusi ekonomi dan penciptaan lapangan kerja jauh lebih penting.	NEGATIF
16	@ferizandra Kalo edukasi dilakukan dengan benar, program vasektomi bisa bantu pengendalian penduduk dan ningkatin kualitas hidup keluarga.	POSITIF
17	Menurut gue keputusan punya anak itu hak pribadi. Pemerintah ga perlu terlalu jauh ngatur sampai dorong masyarakat buat vasektomi.	NEGATIF
18	Kesadaran keluarga berencana itu WAJIB, tp keputusan vasektomi tetap harus jadi pilihan pribadi, bukan tekanan sosial.	NEGATIF
19	Penolakan masyarakat terhadap vasektomi lebih banyak dipengaruhi budaya, stigma, dan pola pikir konservatif yg masih kuat di Jawa Barat.	POSITIF
20	@ipopbase Edukasi yg tepat bisa bantu masyarakat paham manfaat kesehatan reproduksi dan pentingnya pengendalian penduduk secara bertanggung jawab.	POSITIF

Berdasarkan hasil prediksi, model mengklasifikasikan 11 opini (55%) sebagai sentimen negatif dan 9 opini (45%) sebagai sentimen positif, sebagaimana ditunjukkan pada Gambar 8. Hasil tersebut menunjukkan bahwa sentimen negatif masih sedikit mendominasi opini publik terhadap isu vasektomi, sehingga mencerminkan masih adanya pro dan kontra di masyarakat.

Distribusi Sentimen Vasektomi di Media Sosial X Hasil Prediksi SVM Kernel Linear + Skenario 3 (70:30) Terbalik



Gambar 8. Distribusi Hasil Prediksi Sentimen

Hasil prediksi selanjutnya digunakan sebagai dasar visualisasi WordCloud, analisis tematik, dan penyusunan rekomendasi strategis untuk mengidentifikasi pola opini masyarakat terhadap isu vasektomi.

Visualisasi *WordCloud* digunakan untuk mengidentifikasi kata-kata yang paling sering muncul pada kelompok sentimen positif dan negatif hasil prediksi model SVM. Representasi visual ini memberikan gambaran awal mengenai topik yang dominan dibahas masyarakat terkait isu vasektomi. Hasil *WordCloud* untuk masing-masing kelompok sentimen ditunjukkan pada Gambar 9 dan Gambar 10.



Gambar 9. WordCloud Positif



Gambar 10. WordCloud Negatif

Berdasarkan Gambar 9, sentimen positif didominasi oleh kata-kata seperti *edukasi*, *program*, *manfaat*, *keluarga*, dan *kesehatan*, yang menunjukkan bahwa opini positif lebih banyak menyoroti pentingnya edukasi kesehatan reproduksi dan manfaat program vasektomi. Sementara itu, Gambar 10 menunjukkan bahwa sentimen negatif didominasi oleh kata-kata seperti *kemiskinan*, *lapangan kerja*, *ekonomi*, dan *bukan*, yang mengindikasikan bahwa sebagian masyarakat memandang vasektomi bukan sebagai solusi utama terhadap permasalahan sosial dan ekonomi.

Berdasarkan hasil visualisasi *WordCloud*, dilakukan analisis tematik untuk mengidentifikasi pola utama opini masyarakat terhadap isu vasektomi. Proses analisis menghasilkan lima tema utama sebagaimana ditunjukkan pada Tabel 8.

Tabel 8. Hasil Pengkodean Tema Analisis Tematik

Tema	Kode / Kata Kunci	Opini Baris Ke (Tabel 4.20)
Edukasi dan Sosialisasi Kesehatan Reproduksi	Edukasi, Sosialisasi, Kesehatan Reproduksi, Manfaat, Pemahaman, Miskonsepsi	2, 3, 6, 11, 13, 16, 20
Pengendalian Penduduk dan Kesejahteraan Keluarga	Pengendalian Penduduk, Keluarga Berencana, Kualitas Hidup, Kesejahteraan Keluarga, Ledakan Penduduk	2, 3, 10, 16, 20
Kondisi Ekonomi dan Kemiskinan	Kemiskinan, Pengangguran, Ekonomi, Lapangan Kerja, Kesejahteraan, Harga Kebutuhan	1, 7, 9, 14, 15
Budaya, Stigma, dan Pandangan Konservatif	Budaya, Stigma, Konservatif, Kebiri, Pro Kontra, Penerimaan Masyarakat	5, 10, 11, 12, 13, 19
Hak Pribadi dan Penolakan terhadap Vasektomi	Hak Pribadi, Tidak Setuju, Bukan Solusi, Pilihan Pribadi, Tekanan Sosial	4, 8, 17, 18

Hasil analisis menunjukkan bahwa opini masyarakat didominasi oleh lima tema, yaitu edukasi kesehatan reproduksi, pengendalian penduduk dan kesejahteraan keluarga, kondisi

ekonomi, budaya dan stigma, serta hak pribadi. Tema-tema tersebut menjadi dasar dalam penyusunan rekomendasi strategis.

Berdasarkan hasil analisis tematik, disusun rekomendasi strategis sebagai masukan bagi Pemerintah Provinsi Jawa Barat dan BKKBN Jawa Barat dalam meningkatkan efektivitas komunikasi publik mengenai program vasektomi. Rekomendasi strategis berdasarkan setiap tema disajikan pada Tabel 9.

Tabel 9. Rekomendasi Strategis Berdasarkan Hasil Analisis Tematik

Tema	Rekomendasi Strategis
Edukasi dan Sosialisasi Kesehatan Reproduksi	Meningkatkan edukasi mengenai kesehatan reproduksi pria dan prosedur vasektomi melalui media sosial, fasilitas kesehatan, dan penyuluhan masyarakat.
Pengendalian Penduduk dan Kesejahteraan Keluarga	Menyampaikan program vasektomi sebagai bagian dari upaya pengendalian penduduk dan peningkatan kesejahteraan keluarga.
Kondisi Ekonomi dan Kemiskinan	Mengintegrasikan komunikasi program vasektomi dengan kebijakan peningkatan kesejahteraan dan penyediaan lapangan kerja sehingga tidak dipersepsikan sebagai solusi tunggal kemiskinan.
Budaya, Stigma, dan Pandangan Konservatif	Melibatkan tokoh masyarakat, tenaga kesehatan, dan figur publik untuk mengurangi stigma dan miskonsepsi terhadap vasektomi.
Hak Pribadi dan Penolakan terhadap Vasektomi	Menegaskan bahwa vasektomi merupakan pilihan sukarela serta menghormati hak individu dalam mengambil keputusan reproduksi.

5. KESIMPULAN

- Pendekatan pelabelan semi-otomatis yang mengombinasikan Target-Oriented dan Indonesia Sentiment Lexicon (InSet) berhasil menghasilkan 3.024 data berlabel, sehingga mampu menggabungkan ketepatan anotasi manual dengan efisiensi pelabelan otomatis.
- Algoritma Support Vector Machine (SVM) dengan kernel Linear memberikan performa terbaik. Pada skenario 70:30, model memperoleh Balanced Accuracy 74,09%, Macro Average Precision 73,32%, Macro Average Recall 74,09%, dan Macro Average F1-Score 73,64%. Hasil Stratified K-Fold Cross Validation (K=10) juga menunjukkan performa yang konsisten dengan Macro Average F1-Score 72,78%.
- Pengujian terhadap 20 opini terbaru menunjukkan model mampu mengklasifikasikan *unseen data* dengan distribusi 55% sentimen negatif dan 45% sentimen positif. Analisis lanjutan menunjukkan bahwa opini masyarakat didominasi oleh isu edukasi kesehatan reproduksi, pengendalian penduduk, kondisi

ekonomi, budaya dan stigma, serta hak pribadi.

- Kelebihan penelitian ini adalah penerapan pelabelan semi-otomatis yang meningkatkan efisiensi pelabelan tanpa menghilangkan peran anotasi manual sebagai *ground truth*. Namun, pendekatan TF-IDF dan SVM masih terbatas dalam memahami konteks semantik, seperti sarkasme, ironi, dan *code-mixing*.

UCAPAN TERIMA KASIH

Tuhan Yang Maha Esa atas penyertaan-Nya sehingga penelitian ini dapat terselesaikan. Penghargaan yang tulus juga diberikan kepada Universitas Singaperbangsa Karawang, khususnya Fakultas Ilmu Komputer, atas lingkungan akademik yang mendukung proses penelitian ini. Penulis turut berterima kasih kepada dosen pembimbing atas arahan serta umpan balik yang konstruktif dalam penyusunan artikel ilmiah ini.

DAFTAR PUSTAKA

- BPS. (2026). Statistik Indonesia 2026. *Statistik Indonesia* 2026. <https://www.bps.go.id/id/publication/2026/02/27/a43f03f45543dc4e9942f44c/statistik-indonesia-2026.html>
- Sari, A. P., Rahmadini, G., Charlina, H., Ramadan, M. I., & Pradani, Z. E. (2023). Analisis Masalah Kependudukan Di Indonesia. *Journal of Economic Education*, 2(1), 29–37. <https://online-journal.unja.ac.id/JEec/article/view/23180>
- Paramita, Hartati, D., Sumiati, & Sucita, T. (2025). Faktor-Faktor yang Berpengaruh Terhadap Pemilihan Jenis KB MKJP (Metode Kontrasepsi Jangka Panjang) Pada Akseptor KB di UPTD Puskesmas Babulu. *INNOVATIVE: Journal Of Social Science Research*, 5(2), 1–16. <https://j-innovative.org/index.php/Innovative/article/view/18794>
- BKKBN. (2025). *No Title*. <https://kampungkab.bkkbn.go.id/statistik/22/jumlah-peserta-kb-per-mix-kontrasepsi>
- Tempo. (2025). *Dedi Mulyadi Akan Jadikan Vasektomi sebagai Syarat Terima Bansos*. Tempo. <https://www.tempo.co/politik/dedi-mulyadi-akan-jadikan-vasektomi-sebagai-syarat-terima-bansos-1284804>
- Mawaaddah. (2025). Analisis Opini Publik tagar #JanganJadiDosen pada media sosial X (Twitter). *Ikraith-Humaniora*, 9(2), 408–416.

- <https://journals.upi-yai.ac.id/index.php/ikraith-humaniora/article/view/4827>
- [7] DataIndonesia.id. (2025). 8 negara dengan pengguna X terbesar di Dunia. Instagram. <https://www.instagram.com/p/DG8EW-zJWFh/>
- [8] Widyasari, A. P. W., & Muljono, M. (2026). Sentiment Analysis of YouTube Comments on the 2025 DPR RI Demonstration Using Machine Learning. *Journal of Applied Informatics and Computing (JAIC)*, 10(2), 1729–1740. <https://jurnal.polibatam.ac.id/index.php/JAIC/article/view/12254/3527>
- [9] Mulyani, A., Khoerunisa, S., & Kurniadi, D. (2025). Comparison of KNN and SVM Algorithms Performance Using SMOTE to Classify Diabetes. *Jurnal Nasional Teknik Elektro Dan Teknologi Informasi*, 14(1), 25–34. https://www.researchgate.net/publication/391129647_Perbandingan_Kinerja_Algoritma_KNN_dan_SVM_Menggunakan_SMOTE_untuk_Klasifikasi_Penyakit_Diabetes
- [10] Damayanti, N. M., Ariningtyas, I. D., Icham, M. I. A., & Sari, A. P. (2025). Analisis Sentimen Publik Pada Tagar # Btscomeback Di Platform X Menggunakan IndobertTweet. *JITET (Jurnal Informatika Dan Teknik Elektro Terapan)*, 13(3). <https://journal.eng.unila.ac.id/index.php/jitet/article/view/7176/3329>
- [11] Hoover, A. T., Shattuck, D., & Andes, K. L. (2024). Vasectomy provider decision-making balancing autonomy and non-maleficence: qualitative interviews with providers. *Gates Open Research*, 7(132), 1–32. <https://doi.org/10.12688/gatesopenres.15036.2>
- [12] Hidayatullah, R., Abdul Fatah, D., & Yasid, A. (2025). Penerapan Algoritma K-Nearest Neighbor Pada Minat Beli Mobil Bekas Menggunakan Pendekatan Crisp-Dm. *JATI (Jurnal Mahasiswa Teknik Informatika)*, 9(2), 2210–2217. <https://doi.org/10.36040/jati.v9i2.13024>
- [13] Imaddudin, S., Astuti, I., & Ruhama, S. (2025). Studi Sentimen Masyarakat terhadap PSSI di Era Erick Thohir menggunakan Algoritma Support Vector Machine (SVM) pada Media Sosial X. *Jurnal Penelitian Multidisiplin Bangsa*, 1(8), 1003–1013. <https://doi.org/10.59837/jpnmb.v1i8.193>
- [14] Koto, F., & Rahmaningtyas, G. Y. (2017). InSet Lexicon: Evaluation of a Word List for Indonesian Sentiment Analysis in Microblogs. *Proceedings of the 2017 International Conference on Asian Language Processing, IALP 2017*, 391–394. <https://doi.org/10.1109/IALP.2017.8300625>
- [15] Rahmat, Y. S., Renata, S., Belo, A., Dharmawan, W. S., Annisa, R., & Fitriana, L. A. (2026). Perbandingan Algoritma Machine learning Untuk Analisis Sentimen Aplikasi Instagram Berbasis NLP. *Jurnal Komputer Teknologi Informasi Sistem Komputer*, 5(1), 532–543. https://www.researchgate.net/publication/406915478_Perbandingan_Algoritma_Machine_learning_Untuk_Analisis_Sentimen_Aplikasi_Instagram_Berbasis_NLP
- [16] Alaei, A., Wang, Y., Bui, V., & Stantic, B. (2023). Target-Oriented Data Annotation for Emotion and Sentiment Analysis in Tourism Related Social Media Data. *Future Internet*, 15(4). <https://www.mdpi.com/1999-5903/15/4/150>
- [17] Afdal, M., & Elita, L. R. (2022). Penerapan Text Mining Pada Aplikasi Tokopedia Menggunakan Algoritma K-Nearest Neighbor. *Jurnal Ilmiah Rekayasa Dan Manajemen Sistem Informasi*, 8(1), 78–87. <http://dx.doi.org/10.24014/rmsi.v8i1.16595>
- [18] Panjaitan, E. C. S., & Indra, E. (2025). Exploratory Data Analysis Historical Cryptocurrency Ezra. *Jurnal Sistem Informasi Dan Ilmu Komputer Prima (JUSIKOMP)*, 8(2), 90–100. <https://jurnal.unprimdn.ac.id/index.php/jusikom/article/view/6548>
- [19] Salam, R. R., Jamil, M. F., Ibrahim, Y., Rahmadden, R., Soni, S., & Herianto, H. (2023). Sentiment Analysis of Cash Direct Assistance Distribution for Fuel Oil Using Support Vector Machine. *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, 3(1), 27–35. <https://doi.org/10.57152/malcom.v3i1.590>
- [20] Aliyah Salsabila, N., Ardhitto Winatmoko, Y., Akbar Septiandri, A., & Jamal, A. (2018). Colloquial Indonesian Lexicon. *Proceedings of the 2018 International Conference on Asian Language Processing, IALP 2018*, 226–229. <https://doi.org/10.1109/IALP.2018.8629151>
- [21] Hapsari, Y., Mujahidin, S., & Fadhliana, N. (2023). Analisis Sentimen Isu Vaksinasi COVID-19 Pada Twitter Menggunakan Metode Naïve Bayes dan Pembobotan TF-IDF Tokenisasi 1-2. *SPECTA Journal of Technology*, 7(2), 573–583. <https://doi.org/10.35718/specta.v7i2.812>
- [22] Yutika, C. H., Adiwijaya, A., & Faraby, S. Al. (2021). Analisis Sentimen Berbasis Aspek pada Review Female Daily Menggunakan TF-IDF dan Naïve Bayes.

- Jurnal Media Informatika Budidarma*, 5(2), 422–430.
<https://doi.org/10.30865/mib.v5i2.2845>
- [23] Darmayasa, I. N. O., ER, N. A. S., Kadyanan, I. G. A. G. A., & Karyawati, A. A. I. N. E. (2025). Pengaruh Teknik Penanganan Negasi Dalam Analisis Sentimen. *Jurnal Teknologi Informasi Dan Ilmu Komputer (JTIK)*, 12(2), 275–282.
<https://doi.org/10.25126/jtik.2025129079>
- [24] Astriningsih, W. (2023). Identifikasi Multi Aspek dan Sentimen Analisis pada Review Hotel Menggunakan Deep learning. *Universitas Islam Indonesia*.
<https://dspace.uui.ac.id/handle/123456789/46527>
- [25] Baehaqi, F., & Cahyono, N. (2023). Analisis Sentimen Terhadap Cyberbullying Pada Komentar di Instagram Menggunakan Algoritma Naïve Bayes. *Indonesian Journal of Computer Science*, 13(1), 1051–1063.
<https://doi.org/10.33022/ijcs.v13i1.3301>
- [26] Yanti, S. N., & Yuhandri, S. (2024). Implementasi K-Nearest Neighbor Berbasis Particle Swarm Optimization untuk Analisis Sentimen Generasi Sandwich. *Jurnal KomtekInfo*, 11(4), 371–379.
<https://doi.org/10.35134/komtekinfo.v11i4.586>
- [27] Wong, J., & Yoannita, Y. (2026). IMPLEMENTASI TF-IDF DAN COSINE SIMILARITY UNTUK REKOMENDASI RESEP MASAKAN INDONESIA BERDASARKAN BAHAN MAKANAN. *JATI (Jurnal Mahasiswa Teknik Informatika)*, 10(1), 577–584.
<https://ejournal.itn.ac.id/jati/article/view/16816>
- [28] Nurhopipah, A., & Magnolia, C. (2022). Perbandingan Metode Resampling Pada Imbalanced Dataset Untuk Klasifikasi Komentar Program Mbkm. *JUPIKOM: Jurnal Publikasi Ilmu Komputer Dan Multimedia*, 1(2), 9–22.
<https://journalcenter.org/index.php/jupikom/article/view/862>
- [29] Badriyah. (2024). Penerapan Backpropagation Neural Network Dan Myxl Di Google Play Store Oleh : Badriyah. *UNIVERSITAS ISLAM NEGERI MAULANA MALIK IBRAHIM*. <http://etheses.uin-malang.ac.id/72767/>
- [30] Ramlan, R., Satyahadewi, N., & Andani, W. (2023). Analisis Sentimen Pengguna Twitter Menggunakan Support Vector Machine Pada Kasus Kenaikan Harga BBM. *Jambura Journal of Mathematics*, 5(2), 431–445.
<https://doi.org/10.34312/jjom.v5i2.20860>
- [31] Nadhifah, S., Aini, F. N., Kusumawardhani, H. H., & Febrianto, M. Y. F. (2024). Analisis Sentimen Ulasan Pengguna Aplikasi pada Google Play Store Menggunakan Algoritma Support Vector Machine. *Surya Informatika*, 14(1), 1–6.
<https://doi.org/10.48144/suryainformatika.v14i1.1787>
- [32] Padilla, R., Passos, W. L., Dias, T. L. B., Netto, S. L., & Da Silva, E. A. B. (2021). A comparative analysis of object detection metrics with a companion open-source toolkit. *Electronics (Switzerland)*, 10(3), 1–28.
- [33] Prusty, S., Patnaik, S., & Dash, S. K. (2022). SKCV: Stratified K-fold cross-validation on ML classifiers for predicting cervical cancer. *Frontiers in Nanotechnology*, 4(972421), 1–12.
<https://doi.org/10.3389/fnano.2022.972421>
- [34] Devi, D. L., Arifiyanti, A. A., & Wati, S. F. A. (2024). Analisis Sentimen Ulasan Pengguna Access By Kai Menggunakan Metode Word2Vec Dan Algoritma Svm. *Jurnal Informatika Dan Teknik Elektro Terapan*, 12(3).
<https://journal.eng.unila.ac.id/index.php/jitet/article/view/4892/1991>
- [35] Afandy, S. N. (2025). Analisis Sentimen Ulasan Produk Kuliner Lokal Menggunakan Model BERT: Studi Kasus pada UMKM Pawonkoe. *November*, 38–48.
<https://jurnaldrpm.budiluhur.ac.id/index.php/Kresna/article/view/258/190>
- [36] Mubarakah, A. S., Irawan, B., & Prihartono, W. (2025). Penerapan Algoritma Naïve Bayes Untuk Analisis Sentimen Ulasan Pengguna Aplikasi Akademi di Google Play Store. *JITET (Jurnal Informatika Dan Teknik Elektro Terapan)*, 13(3), 2295–2305.
<https://journal.eng.unila.ac.id/index.php/jitet/article/download/6191/3536>
- [37] Gupta, S. C., & Goel, N. (2025). A Weighted-Average Approach Evaluation of Multi-Class Emotion Analysis Model Using Text Mining and Machine Learning. *Cureus Journal of Computer Science*.
<https://doi.org/10.7759/s44389-025-04827-3>
- [38] Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). Crisp-Dm. *SPSS Inc*, 78, 1–78. <http://www.crisp-dm.org/CRISPWP-0800.pdf>