Institut Riset dan Publikasi Indonesia (IRPI)
MALCOM: Indonesian Journal of Machine Learning and Computer Science
Journal Homepage: https://journal.irpi.or.id/index.php/malcom
Vol. 5 Iss. 2 April 2025, pp: 587-597
ISSN(P): 2797-2313 | ISSN(E): 2775-8575

Utilization of Machine Learning in Analyzing Sentiment Towards the TAPERA Program on Digital X Platform

Pemanfaatan Machine Learning dalam Menganalisis Sentimen Terhadap Program TAPERA di Platform Digital X

Aziz Musthafa[1*], Triana Harmini[2], Abid Rafiq[3], Nurhana Marantika[4]

[1,2,3]Program Studi Teknik Informatika, Universitas Darussalam Gontor, Indonesia
[4]Program Studi Ilmu Komunikasi, Universitas Darussalam Gontor, Indonesia

E-Mail: [1]aziz@unida.gontor.ac.id, [2]triana@unida.gontor.ac.id,
[3]abidrafiq@mhs.unida.gontor.ac.id, [4]hana@unida.gontor.ac.id

Received Dec 1st 2024; Revised Feb 12th 2025; Accepted Feb 22th 2025; Available Online Mar 21th 2025, Published Jan 21th 2025
Corresponding Author: Aziz Musthafa

Abstract

Tabungan Perumahan Rakyat (TAPERA) is an Indonesian government program aimed at addressing housing issues for low- and middle-income communities but has received mixed responses due to policy changes. This study analyzes social media user sentiment on Platform X regarding the TAPERA policy to determine whether the dominant sentiment is positive, negative, or neutral. Sentiment analysis was conducted using the Crisp-DM method with Machine Learning algorithms, namely Support Vector Machine (SVM) and Random Forest. A total of 2,936 comments were collected from the Digital X Platform, with the labeling process validated by Communication Science experts to ensure accuracy. Evaluation results using the confusion matrix show that the SVM model outperforms with an accuracy of 88%, compared to 86% for Random Forest. The classification results indicate that the majority of public responses are negative, with 1,874 comments (63.9%). This dominance of negative sentiment reflects public dissatisfaction with the TAPERA program overall. Meanwhile, positive sentiment accounts for only 527 comments (18%), indicating limited appreciation, while neutral sentiment comprises 534 comments (18.1%), highlighting the need for further public awareness efforts. This study is expected to serve as a foundation for recommendations to improve TAPERA management, particularly in terms of service, transparency, and communication.

Keyword: Machine Learning, Random Forest, Platform Digital, Support Vector Machine, TAPERA

Abstrak

Tabungan Perumahan Rakyat (TAPERA) adalah program pemerintah Indonesia yang bertujuan mengatasi masalah perumahan bagi masyarakat berpenghasilan rendah dan menengah, namun mendapat beragam respons akibat perubahan kebijakan. Tujuan Penelitian ini menganalisis sentimen pengguna media sosial X terhadap kebijakan TAPERA, untuk mengidentifikasi apakah sentimen yang dominan adalah positif, negatif, atau netral. Metode yang dimanfaatkan untuk menganilisis sentimen yaitu Crisp-DM dengan memanfaatkan model *Support Vector Machine* (SVM) dan *Random Forest* yang merupakan algoritma dari *Machine Learning*. Data dikumpulkan dari Platform Digital X dengan total 2.936 komentar, dan proses pelabelannya divalidasi oleh ahli Ilmu Komunikasi guna memastikan akurasi serta menghindari kesalahan. Hasil evaluasi menggunakan *confusion matrix* dari Model SVM menunjukkan keunggulan dengan akurasi 88%, dibandingkan dengan Random Forest yang memiliki akurasi 86%. Sedangkan hasil klasifikasi model, masyarakat lebih cenderung memberikan respons negatif terhadap perubahan kebijakan program TAPERA yaitu 1.874 komentar (63,9%). Dominasi sentimen negatif ini mencerminkan ketidakpuasan masyarakat terhadap program TAPERA secara umum. Sentimen positif sejumlah 527 komentar (18%), menunjukkan apresiasi terhadap inisiatif pemerintah masih terbatas. Serta Sentimen netral sejumlah 534 komentar (18,1%), menunjukkan kebutuhan informasi untuk meningkatkan pemahaman masyarakat. Penelitian ini diharapkan dapat menjadi pendukung rekomendasi untuk perbaikan pengelolaan TAPERA yang lebih baik, terutama dalam aspek layanan, transparansi, dan komunikasi.

Kata Kunci: Machine Learning, Random Forest, Platform Digital, Support Vector Machine, TAPERA

1. PENDAHULUAN

Tabungan Perumahan Rakyat (TAPERA) adalah gagasan yang diluncurkan oleh Institusi negara Indonesia mengatasi masalah perumahan bagi masyarakat berpenghasilan rendah dan menengah [1]. Program ini bertujuan memberikan akses terhadap kepemilikan hunian yang memadai dan ekonomis dengan menggunakan skema simpan pinjam yang terstruktur. Pada tahun 2024, pemerintah melakukan revisi terhadap Peraturan Pemerintah (PP) Nomor 25 Tahun 2020 terkait Pelaksanaan TAPERA melalui dikeluarkannya PP Nomor 21 Tahun 2024. Berdasarkan PP Nomor 21 Tahun 2024, iuran TAPERA sejumlah 3% dari pendapatan atau penghasilan, dengan perincian 0,5% dibebankan kepada pemberi kerja dan 2,5% kepada karyawan untuk peserta pekerja, sedangkan peserta pekerja mandiri menanggung sendiri 3% [2].

Meskipun program ini bertujuan untuk mengatasi masalah perumahan di Indonesia, namun karena ada perubahan tersebut respons masyarakat terhadap kebijakan ini menuai berbagai tantangan dan perdebatan serta menimbulkan polemik [3]. Sebagian masyarakat menyambut baik program tersebut, sementara sebagian lainnya menyatakan ketidakpuasan dan kekhawatiran terkait efektivitas dan manfaatnya. Dengan berkembangnya teknologi digital, masyarakat kini dapat dengan mudah mengungkapkan opini mereka seperti di berbagai media sosial. Sentimen yang tersebar pada berbagai platform digital salah satunya platform digital X ini dapat mencerminkan persepsi publik terhadap kebijakan TAPERA. Platform digital X dipilih karena fitur *information sharing*-nya menempati peringkat keenam sebagai alasan utama yang mendorong pengguna untuk menggunakan platform tersebut [4].

Untuk memahami sentimen dari opini masyarakat, diperlukan penggunaan analisis sentimen. Analisis sentimen adalah teknik atau metode yang dimanfaatkan sebagai identifikasi emosi yang terkandung dalam tulisan dan mengelompokkan perasaan tersebut menjadi positif , netral atau negatif [5]. Oleh karena itu, dibutuhkan pendekatan yang lebih efektif dan maksimal, salah satunya ialah menggunakan *Machine Learning* untuk menganalisis sentimen publik secara otomatis melalui platform digital. Teknologi *Machine Learning* memberikan solusi yang efektif untuk menganalisis data sentimen yang jumlahnya sangat besar dan kompleks. *Machine Learning* atau pembelajaran mesin adalah cabang kecerdasan buatan yang terus berkembang, memusatkan perhatian pada konsep identifikasi pola dan pembelajaran berbasis komputer [6]. Teknologi ini memanfaatkan algoritma pembelajaran, baik terawasi maupun tidak terawasi, untuk melakukan klasifikasi dan mendukung pengambilan keputusan secara otomatis berdasarkan himpunan data. Dengan mengaplikasikan algoritma ML, data yang bersumber pada platform digital dapat diproses untuk mengidentifikasi pola sentimen positif, negatif, atau netral terhadap TAPERA. Diharapkan pendekatan ini mampu memberikan gambaran yang lebih tepat dan langsung terkait dengan penerimaan masyarakat terhadap program tersebut.

Beberapa penelitian sebelumnya telah menganalisis sentimen terkait program TAPERA dengan menerapkan metode klasifikasi berbasis ML. Penelitian pertama melakukan analisis sentimen program TAPERA menggunakan algoritma K-Nearest Neighbor (K-NN) namun kekurangan dalam penelitian ini masih belum menghasilkan nilai akurasi cukup tinggi pada tahap evaluasi yaitu 62,75%[7]. Penelitian kedua mengenai analisis sentimen program TAPERA menggunakan perbandingan algoritma SVM, Naïve Bayes, dan K-NN. Hasil penelitian menunjukkan akurasi algoritma Naïve Bayes 71,15%, akurasi algoritma K-NN 73,07%, dan akurasi tertinggi yaitu Algoritma SVM 81,73% [8]. Hasil penelitian ini memaparkan bahwa dalam analisis sentimen program TAPERA paling bagus menggunakan algoritma SVM, namun data yang digunakan pada penelitian ini sebesar 519 komentar sehingga perlu pengujian pada data skala yang lebih besar. Serta terdapat kelemahan pada semua pengujian menunjukkan kegagalan prediktif untuk data tweet berlabel netral yang dianggap positif.

Penelitian ketiga analisis sentimen program TAPERA dengan *Naïve Bayes Classifier* dan *Support Vector Machine* (SVM). Hasil penelitian menghasilkan nilai akurasi untuk algoritma Naïve Bayes sebesar 81% sedangkan akurasi algoritma SVM lebih besar yaitu 84%. Penelitian ini menggunakan data 1280 komentar dan mendapatkan akurasi lebih tinggi dibandingkan penelitian kedua sehingga pada penelitian ini menggunakan data penelitian yang lebih besar 2.936 komentar, serta saran di penelitian ketiga agar mendapatkan performa yang lebih baik dapat menggunakan algoritma lainnya seperti *Random Forrest* ini yang melatarbelakangi pada penelitian ini menggunakan *Random Forrest* sebagai pembanding [9].

Penelitian keempat melakukan analisis sentimen pada aplikasi shopee dengan *Random Forrest* menghasilkan nilai akurasi 84,9% [10]. Serta penelitian kelima melakukan analisis sentimen kenaikan harga BBM di Indonesia menggunakan *Random Forrest* menghasilkan akurasi 85,15% [11]. Pada kedua penelitian ini nilai akurasi yang dihasilkan cukup bagus sehingga layak untuk digunakan pembanding pada penelitian kali ini.

Dari penelitian terdahulu didapatkan informasi bahwa algoritma SVM lebih baik dari algoritma lainnya namun perlu membandingkan dengan algoritma *Random Forrest.* Dengan demikian, penelitian ini akan melakukan analisis sentimen terhadap program TAPERA dengan memanfaatkan algoritma SVM dan *Random Forest*. *Support Vector Machine* (SVM) merupakan algoritma dalam pembelajaran terawasi yang terkenal untuk tugas klasifikasi dan regresi, baik pada data linier maupun non-linier. Inti dari SVM adalah menentukan hyperplane yang memaksimalkan jarak antara kelas-kelas untuk memastikan pemisahan yang

paling efektif. Dalam dua dimensi, hyperplane berbentuk garis, sementara dalam tiga dimensi berupa bidang, dan pada dimensi yang lebih tinggi tetap disebut hyperplane. Secara keseluruhan, SVM berfokus pada menemukan pemisah terbaik untuk mengklasifikasikan data secara akurat [12]. Sedangkan Algoritma *Random Forest* merupakan pendekatan yang memanfaatkan pohon keputusan sebagai pengklasifikasi dasar, yang dibangun dan digabungkan melalui proses pengambilan sampel terintegrasi untuk menghasilkan model prediksi yang lebih kokoh dan tepat [13].

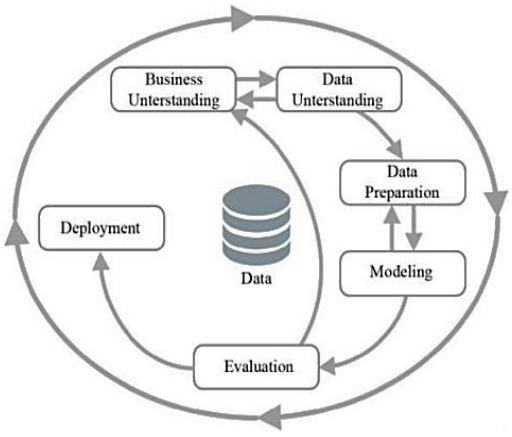
Pendekatan ini diharapkan mampu menghasilkan algoritma terbaik yang lebih sesuai untuk analisis sentimen program TAPERA. Selain itu, penelitian ini diharapkan menjadi acuan penting dalam mengevaluasi efektivitas analisis sentimen yang diterapkan pada program tersebut. Hasil penelitian ini juga berpotensi memberikan rekomendasi yang bermanfaat untuk pengembangan dan peningkatan program TAPERA di masa depan, sekaligus menjadi dasar yang kuat bagi penelitian lanjutan.

2. BAHAN DAN METODOLOGI PENELITIAN

Tujuan Penelitian ini menganalisis sentimen pengguna media sosial X terhadap kebijakan TAPERA, untuk mengidentifikasi apakah sentimen yang dominan adalah positif, negatif, atau netral. Data yang didapat sejumlah 2.936 data akhir dalam rentang waktu 21 Juni hingga 23 Juli 2024. Mengenai algoritma yang digunakan yaitu SVM dan *Random Forrest* dimana pemilihan model didasarkan pada penelitian terdahulu terkait analisis sentimen pada program TAPERA dimana SVM yang terbaik sesuai tingkat akurasi, namun terdapat saran agar dibandingkan dengan algoritma *Random Forrest* serta penambahan data penelitian. Dari penelitian terdahulu juga mengenai analisis sentimen, *Random Forrest* memiliki tingkat akurasi baik sehingga layak untuk digunakan sebagai pembanding. Penelitian ini menggunakan Metode CRISP-DM sebagai langkah-langkah proses penelitian.

Metode CRISP-DM (*Cross-Industry Standard Process Model for Data Mining*) adalah Metodologi penambangan data yang dikembangkan oleh kelompok perusahaan di bawah bimbingan Komisi Eropa pada tahun 1996. Metode ini dirancang sebagai standar untuk menerapkan proses data mining dalam berbagai konteks. Tujuan utamanya adalah memberikan kerangka kerja sistematis untuk menganalisis strategi yang digunakan dalam memecahkan masalah penelitian maupun tantangan bisnis. CRISP-DM dibagi menjadi enam langkah utama: Pemahan dari Segi Bisnis yaitu Mengidentifikasi tujuan bisnis dan merencanakan langkah strategis. *Data Understanding* yaitu Mengumpulkan, mendeskripsikan, dan mengeksplorasi data. *Data Preparation* yaitu Membersihkan data dari nilai kosong atau tidak relevan, mengintegrasikan, dan memformat data untuk analisis. *Modeling* yaitu Memilih teknik pemodelan, merancang pengujian, membangun model, dan mengevaluasi performanya. *Evaluation* yaitu Mengevaluasi hasil model dengan mempertimbangkan tujuan bisnis dan menvisualisasikannya. *Deployment* yaitu Merekomendasikan strategi dan langkah untuk implementasi lebih lanjut [14].

Metodologi ini menyediakan panduan yang terstruktur untuk memastikan keberhasilan dalam proyek data mining, mulai dari pemahaman kebutuhan bisnis hingga implementasi solusi berbasis data. Gambar 1 menunjukkan alur metode CRISP-DM yang diterapkan pada penelitian analisis sentimen program TAPERA.



Gambar 1. Alur Proses CRISP-DM [15]

2.1. *Business Understanding*

Pada tahapan *Business understanding* hal pertama yang dilakukan adalah mengidentifikasi tujuan bisnis. Fokus utama dari penelitian ini adalah untuk mengeksplorasi pandangan masyarakat mengenai program TAPERA berdasarkan media sosial X. Walaupun program ini dirancang untuk mengatasi masalah perumahan di Indonesia, perubahan yang terjadi memunculkan beragam respons dari masyarakat. Sebagian mendukung program tersebut, sementara yang lain merasa kurang puas dan khawatir terhadap efektivitas

serta manfaatnya. Dengan bantuan ML yang menggunakan algoritma SVM dan *Random Forrest*, Peneliti dapat memahami secara umum bagaimana persepsi masyarakat terhadap program ini, apakah lebih mengarah pada pandangan positif, netral, atau negatif.

2.2. *Data Understanding*

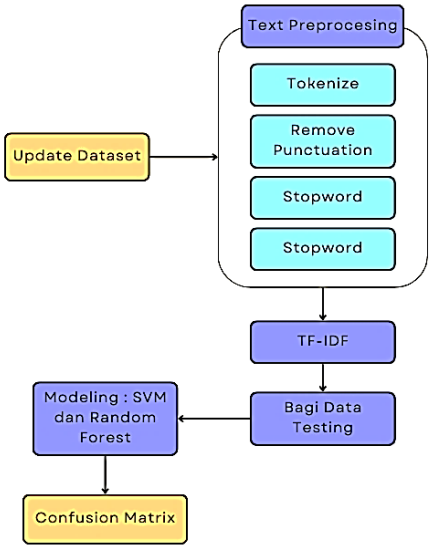
Pada tahap ini, pengumpulan data dilakukan melalui teknik crawling di platform media sosial X dengan menggunakan bahasa pemrograman *Python*. Pengumpulan data dilakukan menggunakan *Tweet-harvest* untuk melakukan *crawling* dari media social X. Proses ini menggunakan kata kunci "TAPERA lang:id" dan menghasilkan 5.011 data mentah yang dikumpulkan dalam rentang waktu 21 Juni hingga 23 Juli 2024. Data yang diperoleh kemudian melalui tahap pembersihan, termasuk penghapusan duplikat dan eliminasi data yang tidak relevan secara manual, menghasilkan 2.936 data akhir. Selanjutnya, proses pelabelan data divalidasi oleh dosen ahli di bidang Ilmu Komunikasi yaitu Nurhana Marantika, S.Sos.I., M.A. untuk memastikan akurasi dan menghindari kesalahan dalam proses pelabelan.

2.3. *Data Preparation*

Tahap Persiapan Data adalah langkah awal untuk mempersiapkan data yang melibatkan berbagai aktivitas dalam membangun dataset final. Pada tahap ini, dilakukan beberapa pemrosesan awal teks, seperti Tokenisasi, Penghapusan Tanda Baca, Penghapusan Stopword, dan Stemming. Langkah pertama, yaitu Tokenisasi, bertujuan dalam membagi suatu teks ke dalam kata yang membentuknya. Normalisasi Teks bertujuan mengganti kata yang tidak jelas kedalam bentuk lebih jelas agar lebih sederha dipahami. Proses berikutnya adalah menghaspus *stopword*, yang bertujuan dalam menghilangkan kata-kata umum yang kurang memiliki pengaruh signifikan dalam kalimat. Di tahap akhir, Stemming dilakukan dalam mengganti kata berimbuhan kedalam bentuk dasarnya. Setelah itu, untuk mempersiapkan data agar dapat diproses pada tahap model, diperlukan Pemberian Bobot Kata (TF-IDF), yang melibatkan langkah pertama menghitung Term Frequency (TF), kemudian menentukan Inverse Document Frequency (IDF), dan terakhir, menggabungkan keduanya untuk memperoleh bobot akhir setiap kata [16].

2.4. *Modeling*

Pada tahap ini, data dipisahkan jadi 2 bagian, yaitu data pembelajaran (training data) dan data ujicoba (testing data). Data pembelajaran dikerjakan agar dapat mempelajari pola serta karakteristik nilai, sementara data ujicoba berfungsi dalam menilai kinerja algoritma yang telah dilatih. Pemisahan data diperlukan dengan proporsi 80% pada data pembelajaran dan 20% pada data ujicoba. tahapan pembuatan model bisa dipaparkan di Gambar 2.



Gambar 2. Alur Proses *Modeling*

Sesuai dengan Gambar 2, data yang sudah dibersihkan ini kemudian diubah menjadi vektor dengan menggunakan TF-IDF untuk diproses sebagai input dalam algoritma pembelajaran mesin. Dari seluruh data yang ada, data tersebut dipisahkan menjadi dua bagian untuk keperluan pelatihan dan pengujian dengan proporsi 80:20, di mana 80% digunakan untuk pelatihan dan 20% untuk pengujian. Selanjutnya, dilakukan pemodelan menggunakan algoritma SVM dan Random Forest.

SVM menerapkan fungsi hipotesis linier dalam ruang fitur dengan dimensi tinggi dan menggunakan bias pembelajaran yang didasarkan pada prinsip statistik. Algoritma ini memisahkan dataset menjadi dua

kelompok yang dibedakan oleh sebuah hyperplane, dengan kelas pertama diberi label 1 dan kelas kedua diberi label -1.

$$X_i.W + b \geq 1 \; untuk \; Y_i = 1 \tag{1}$$

$$X_i.W + b \leq 1 \; untuk \; Y_i = -1 \tag{2}$$

Persamaan dalam menghitung hasil klasifikasi menggunakan *hyperplane* satuan, berdasarkan nilai b serta w yang telah ditentukan, dapat dirumuskan seperti persamaan 3 dan 4:

$$F((x)) = sign\big(w.\emptyset(x)\big) + b \tag{3}$$

$$F((x)) = sign\left(\sum_{i=1}^{n} a_i \, y_i \, \emptyset(x_i)^T.\emptyset(x) + b\right) \tag{4}$$

Pada persamaan 3 dan 4 terdapat keterangan yaitu x_i merupakan nilai dokumen ke i, w yaitu nilai bobot support vector yang ortogonal terhadap *hyperplane*, b yaitu data penyimpangan, serta y_i adalah grup data ke-i. Menetapkan nilai optimal untuk hyperplane dua kategori dengan menggunakan persamaan 5 [17].

$$minimize \int 1[w] = \frac{1}{2}||x||^2 \tag{5}$$

Tahap terakhir pengklasifikasian menggunakan algoritma *Random Forest* dapat ditentukan dengan memanfaatkan angka *Gini Index* dalam menetapkan variabel pemisah yang akan digunakan sebagai akar (root) atau simpul (node). Penghitungan *Gini Index* dan *Gini Split* untuk kategorisasi diselesaikan berdasarkan rumus (6) dan (7).

$$Gini\ Index\ (Si) = 1 - \sum_{i=1}^{C} \pi_i^2 \tag{6}$$

Di mana πi\pi_iπi adalah probabilitas yang terkait dengan *Gini Index* pada bagian i, dan c merepresentasikan jumlah skor fasilitas yang terakumulasi dalam suatu bagian berdasarkan data fasilitas.

$$Gini\ split\ = \sum_{i=1}^{C} \left(\frac{n_i}{n}\right) x\ Gini\ Index\ (Si) \tag{7}$$

Di mana $n_i$merupakan jumlah sampel setelah proses split, n adalah total jumlah sampel pada node tertentu, dan c adalah jumlah skor fasilitas yang termasuk dalam suatu bagian berdasarkan data fasilitas [18].

Kedua algoritma digunakan untuk membandingan akurasi model. Setelah proses pemodelan, tahap berikutnya adalah pengukuran performa klasifikasi dengan *Confusion Matrix.*

2.5. *Evaluation*

Tahap Evaluation adalah proses mengevaluasi kinerja model dengan menggunakan *Confusion Matrix* untuk mengukur performa klasifikasi SVM maupun *random forest*. *Confusion Matrix* merupakan Gambar 3 yang berfungsi untuk menilai tingkat keakuratan prediksi model.

Nilai Prediksi	Nilai Aktual: Positive	Nilai Aktual: Negative
Positive	TP	FP
Negative	FN	TN

Gambar 3. *Confusion Matix*

Seperti yang di tunjukan pada Gambar 3, Pada *confusion matriks* terdapat empat istilah utama. Diantaranya *True Positive* (TP) yaitu jumlah angka aktual positif yang dilabeli benar dari classifer, *True Negative* (TN) adalah keseluruhan nilai aktual negative yang dilabeli benar dari classifer, *False Positive* (FP)

adalah keseluruhan nilai aktual negative yang salah dilabeli oleh classifer, dan *False Negative* (FN) adalah keseluruhan nilai aktual positif bernilai salah dilabeli oleh classifer.

Melalui proses klasifikasi didapatkan hasil Akurasi yang merupakan hasil jumlah rasio prediksi Sesuai dengan jumlah keseluruhan prediksi yang ada sesuai rumus 8 [19].

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \quad (8)$$

2.6. *Deployment*

Tahap *Deployment* merupakan tahap akhir dalam penelitian ini, yang berfokus pada visualisasi data. Visualisasi ini bertujuan untuk menunjukkan kata yang paling sering muncul berdasarkan kategorisasi sentimennya. Dengan visualisasi tersebut, analisis terhadap opini masyarakat yang disampaikan melalui media sosial X menjadi lebih mudah, tanpa harus membaca seluruh komentar secara individu..

3. HASIL DAN PEMBAHASAN

Pada tahap ini dataset yang digunakan dalam penelitian ini berjumlah 2.936, dengan tiga atribut yang dijadikan label klasifikasi, yaitu negatif, netral, dan positif. Setelah data diberi label dan divalidasi oleh ahli yaitu dosen ilmu komunikasi selanjutnya data dibersihkan melalui tahap *data preparation* yang terdiri dari :

3.1. *Casefolding* dan *Tokenize*

Pada tahap ini data berupa kalimat Panjang diperkecil dengan Casefolding dan dipisahkan perkata menggunakan *Tokenize*. Selain itu untuk memaksimalkan proses klasifikasi dilakukan penghapusan hypelink, koma, angka, dan simbol. Tabel 1 menunjukkan hasil perubahan dari kalimat Panjang menjadi hasil setelah *casefolding* dan *tokenize*.

Tabel 1. Hasil *Casefolding* dan *Tokenize*

No	full_text	label	casefolding	tokenizing
1	@Heraloebss @Andiarief__ @pln_123 lu dapet apa...	negatif	lu dapet apa mis dapet TAPERA	[lu, dapet, apa, mis, dapet, TAPERA]
2	@okezonenews Lho katanya TAPERA nggak wajib	netral	lho katanya TAPERA nggak wajib	[lho, katanya, TAPERA, nggak, wajib]

3.2. Normalisasi Teks

Pada tahapan normalisasi, kata baik itu kata kurang baku, singkatan, asing akan dirubah menjadi kata baku yang lebih jelas dimana data kata normalisasi telah disiapkan dalam dataset. Tabel 2 menunjukkan contoh perubahan kata tidak normal menjadi kata normal.

Tabel 2. Hasil Normalisasi Teks

Kata Belum Normal	Kata Normal
kek	seperti
nggak	tidak
lu	Kamu

3.3. *Filtering (Stopword Removal)*

Langkah berikutnya adalah langkah Stopword Removal, yaitu proses yang bertujuan untuk menghapus kata-kata umum dan sering digunakan tetapi tidak memberikan pengaruh signifikan dalam sebuah kalimat. Hasil proses Filtering atau penghilangan data yang kurang berguna untuk proses tahapan berikutnya terdapat pada Tabel 3.

Tabel 3. Luaran proses *Stopword Removal*

No	full_text	label	normalize	stopwords
1	@AndyHuskyyy Dasinya kek TAPERA.. mencekik	negatif	dasinya seperti TAPERA mencekik	dasinya TAPERA mencekik
2	Iuran Wajib TAPERA dari Potong Gaji Pekerja Sw...	netral	iuran wajib TAPERA dari potong gaji pekerja	iuran wajib TAPERA potong gaji pekerja

3.4. *Stemming*

Proses terakhir, yaitu tahap Stemming, bertujuan untuk mengubah kata berimbuhan menjadi bentuk dasarnya. Pada tahap ini, digunakan pustaka Python Sastrawi untuk menghilangkan imbuhan dalam kata-kata berbahasa Indonesia, sehingga menghasilkan bentuk dasar. Hasil dari proses Stemming dapat dilihat pada Tabel 4.

Tabel 4. Hasil proses *Stemming*

No	full_text	label	stopwords	stemming
1	@AndyHuskyyy Dasinya kek TAPERA.. mencekik	negatif	dasinya TAPERA mencekik	dasi TAPERA cekik
2	Iuran Wajib TAPERA dari Potong Gaji Pekerja Sw...	netral	iuran wajib TAPERA potong gaji pekerja swasta	iur wajib TAPERA potong gaji kerja swasta

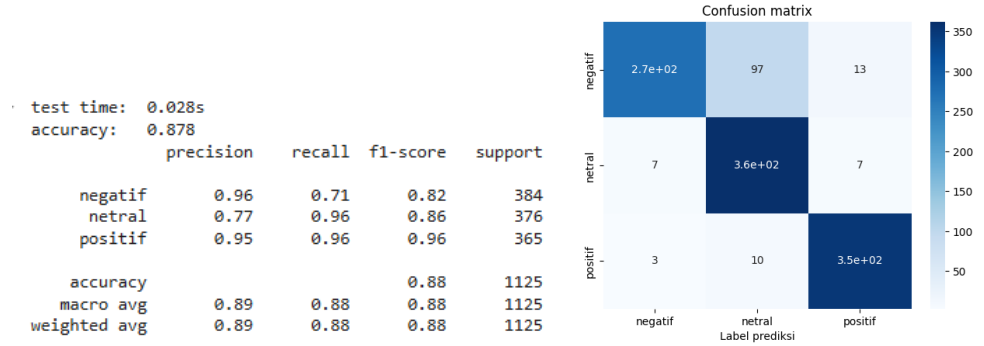
3.5. Pembuatan Model

Dalam proses ini, data dipisahkan menjadi dua kelompok: data pelatihan dan data pengujian. Data latih, yang mencakup 80% dari total dataset yang telah diproses yaitu 2.349, digunakan untuk mempelajari karakteristik data yang benar dan salah guna membangun model klasifikasi. Sementara itu, data uji, yang terdiri dari 20% dari total dataset yaitu 578, digunakan untuk mengevaluasi performa model yang sudah dilatih dengan data tersebut. Proses ini bertujuan untuk mengevaluasi keakuratan dan efektivitas model yang dikembangkan. Dari data pembagian data tersebut akan diimplementasikan kedalam algoritma SVM dan *Random Forest* menggunakan *python.*

3.6. *Evaluation*

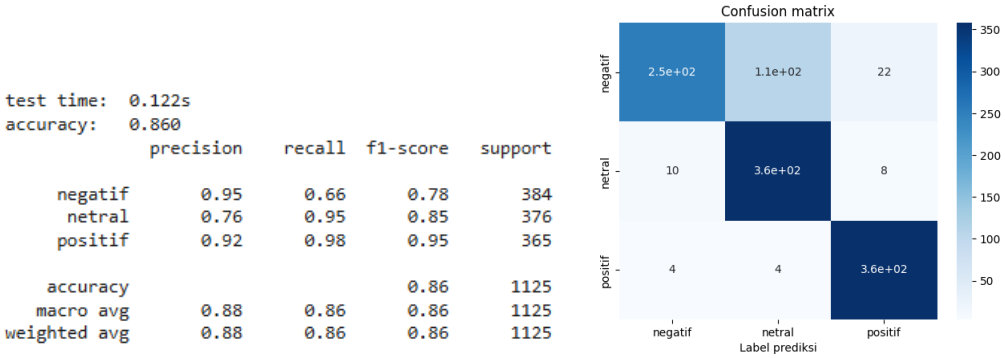
Evaluasi dilakukan dengan memanfaatkan *Confusion Matrix* untuk menghitung tingkat akurasi dari algoritma SVM serta *Random Forest*. Hasil evaluasi ini kemudian dibandingkan untuk menentukan algoritma yang paling efektif dalam analisis sentimen terkait program TAPERA.

1. SVM : dari hasil evaluasi memanfaatkan *Confusion Matrix* untuk memperoleh nilai akurasi 0,869 atau 88% dimana hasil lengkap tertera pada Gambar 4.



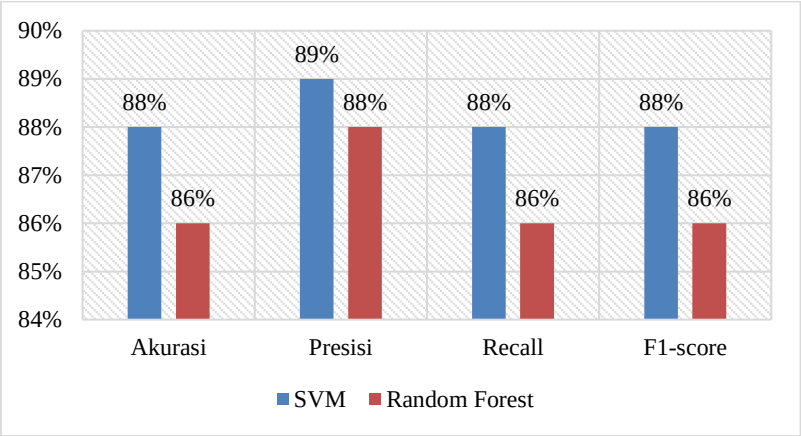
Gambar 4. Hasil *Confusion Matrix* pada Algoritma SVM

2. *Random Forest* : dari hasil evaluasi memanfaatkan *Confusion Matrix* memperoleh angka akurasi 0,860 atau 86% dimana hasil lengkap tertera pada Gambar 5.



Gambar 5. Hasil *Confusion Matrix* pada Algoritma *Random Forest*

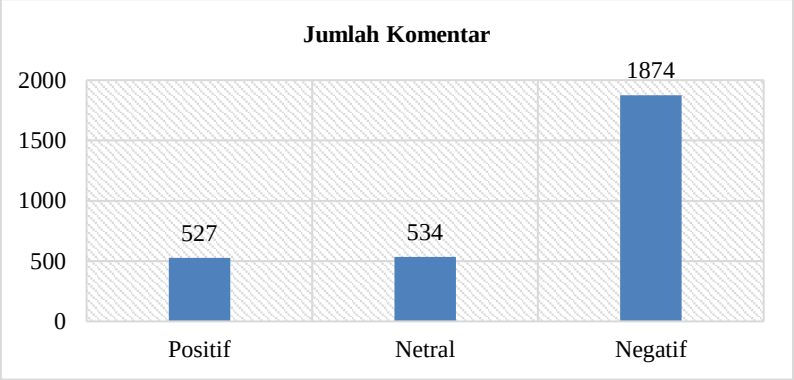
Berdasarkan hasil evaluasi, baik model SVM maupun *Random Forest* menunjukkan kinerja yang cukup baik dalam mengklasifikasikan sentimen terkait Program TAPERA pada dataset media sosial X. Model SVM mencatatkan akurasi sebesar 88%, presisi 89%, recall 88%, dan skor F1 88%. Di sisi lain, model *Random Forest* mencatatkan akurasi 86%, presisi 88%, recall 86%, dan skor F1 86%. Berdasarkan perbandingan tersebut, model SVM menunjukkan kinerja yang sedikit lebih tinggi dibandingkan dengan model *Random Forest*, seperti yang terlihat pada grafik di Gambar. 6.



Gambar 6. Grafik Perbandingan algoritma SVM dan *Random Forest*

3.7. *Deployment*

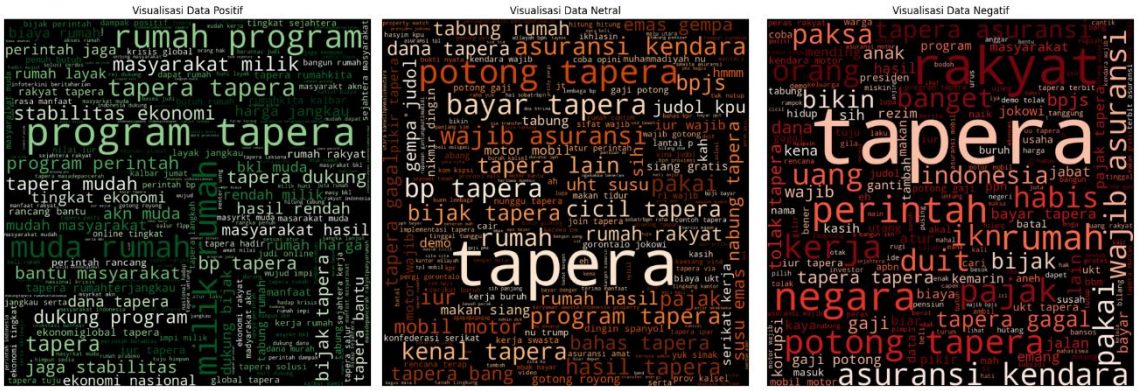
Visualisasi data yang dapat dihasilkan dari *crawling* dari media social X menggunakan kata kunci "TAPERA lang:id" dalam rentang waktu 21 Juni hingga 23 Juli 2024 menghasilkan 2.936 data akhir. Dari data tersebut hasil analisis sentimen program TAPERA berdasarkan label positif sebanyak 527 komentar (18%), netral sebanyak 534 komentar (18,1%), dan negatif menghasil nilai yang relatif banyak yaitu 1.874 komentar (63,9%) yang terdapat pada Gambar 7. Walaupun hasil klasifikasi banyak yang negatif, namun memang perlu dikaji kembali terkait sentiment masyarakat pada beberapa tahun setelah pelaksanaan atau pasca program TAPERA ini.



Gambar 7. Grafik analisis sentiment berdasrkan label

Frekuensi kemunculan kata-kata digambarkan melalui *wordcloud* pada Gambar 8, dapat menggambarkan analisis sentimen lebih dalam sesuai dengan label kategorinya. Seperti yang terlihat pada Gambar 8, *Wordcloud* adalah metode visualisasi teks yang menggambarkan kata yang terlalu sering muncul dalam teks dengan gambaran yang lebih besar atau lebih mencolok, sementara kata yang tidak sering muncul digambarkan dengan gambaran yang lebih kecil [20]. Dengan menggunakan *Wordcloud* dapat melihat kata-kata yang paling banyak muncul di dalam tweet terkait program TAPERA. Kata-kata dengan ukuran lebih besar menunjukkan frekuensi kemunculan lebih besar. Visualisasi ini memudahkan dalam mengenali topik atau permasalahan yang paling sering dibicarakan dalam tweet atau komentar.

Berdasarkan wordcloud dengan kategori positif pada Gambar 8, kata yang sering muncul seperti "program', "rumah","dukung", "milik", dan "masyarakat". Gambaran ini menunjukkan bahwa masyarakat mengapresiasi program TAPERA dalam hal kepemilikan rumah untuk rakyatnya. Namun jumlahnya masih terbatas dan lebih banyak suara yang negatif. Sedangkan pada kategori netral kata yang sering muncul yaitu "TAPERA","potong","asuransi","rumah",dan "kenal". Kata tersebut menunjukkan ungkapan deskriptif tanpa luapan emosi yang mengacu pada implementasi program TAPERA. Hal ini menunjukkan bahwa perlunya kebutuhan informasi untuk meningkatkan pemahaman masyarakat terkait program TAPERA. Kemudian pada kategori negatif kata yang sering muncul yaitu "TAPERA","potong","perintah","gagal","paksa",dan "rakyat". Hal tersebut menunjukkan ketidakpuasan masyarakat terhadap program TAPERA yang terpaksa harus diterima oleh masyarakat.



Gambar 8. Wordcloud analisis sentimen program TAPERA

Hasil penelitian menunjukkan bahwa analisis sentimen terhadap program TAPERA melalui platform X memberikan gambaran persepsi masyarakat yang terbagi ke dalam tiga kategori: positif, netral, dan negatif. Dalam proses klasifikasi, model SVM menunjukkan kinerja lebih baik dengan akurasi 88%, dibandingkan dengan Random Forest yang memiliki akurasi 86%. Berdasarkan pada penelitian ketiga terdahulu dimana hasil evaluasi SVM dan Naïve bayes juga menunjukkan hasil SVM lebih baik [9]. Kemudian saran penelitian tersebut untuk membandingkan dengan Random Forest mengonfirmasi bahwa model SVM masih lebih efektif dalam mengklasifikasikan sentimen masyarakat.

Berdasarkan hasil klasifikasi menggunakan model SVM, dari total 2.936 komentar yang dianalisis, sebanyak 1.874 komentar (63,9%) bersentimen negatif. Dominasi sentimen negatif ini mencerminkan ketidakpuasan masyarakat terhadap program TAPERA secara umum. Analisis Wordcloud pada Gambar 8 menunjukkan bahwa keresahan publik terutama terkait kebijakan pemotongan gaji dalam program TAPERA. Berikut komentar terkait sentiment negatif yang muncul "*Pembahasan TAPERA dipotong 3% dari gaji. Aduh tolong dikasi opsi dapat diambil atau tidak biarkan karyawan yg memilih. Ini kasihan banget yg gajinya di bawah Rp 10 jt. Bahkan yg di atas Rp 10jt juga banyak banget potongannya Kalau yg ini tdk setuju*" dan "*Ini tuh Salah 1 akibat uang belanja kurang.. Penghasilan suami smakin kurang krna banyak potongannya.. Pajak token listrik air bpjs skolah anak sandang pangan TAPERA Dan sebentar lagi akan ada ASUransi.. Suram dah klo bgni terus..*".

Di sisi lain, sentimen positif hanya ditemukan dalam 527 komentar (18%), jumlah yang jauh lebih sedikit dibandingkan dengan sentimen negatif. Hal ini menunjukkan bahwa meskipun program TAPERA bertujuan untuk menyediakan perumahan bagi masyarakat berpenghasilan rendah, namun apresiasi terhadap inisiatif pemerintah masih terbatas. Serta Sentimen netral sejumlah 534 komentar (18,1%), menunjukkan kebutuhan informasi untuk meningkatkan pemahaman masyarakat terkait program TAPERA.

Banyaknya kritik negatif menunjukkan bahwa implementasi program TAPERA belum sepenuhnya memenuhi ekspektasi masyarakat. Hal ini menunjukkan kejelasan antara hasil penelitian ini dengan studi kasus TAPERA. Temuan ini dapat menjadi dasar rekomendasi untuk perbaikan pengelolaan TAPERA, terutama dalam aspek layanan, transparansi, dan komunikasi. Diperlukan strategi komunikasi yang lebih efektif guna meredam sentimen negatif serta meningkatkan pemahaman masyarakat mengenai manfaat program ini. Namun data yang di gunakan dalam penelitian ini masih 1 platform yaitu media sosial X, maka perlunya dilanjutkan dalam penelitian selanjutnya dengan multi platform media sosial.

4. KESIMPULAN

Berdasarkan hasil evaluasi dan analisis sentimen terhadap Program TAPERA melalui model SVM dan Random Forest, Model SVM menunjukkan sedikit keunggulan dengan akurasi 88%, presisi 89%, recal 88%, serta F1-score 88%. Di pihak lain, algoritma *Random Forest* juga Menampilkan performa yang unggul dengan akurasi 86%, presisi 88%, recal 86%, serta F1-score 86%, meskipun sedikit lebih rendah dibandingkan dengan model SVM. Kemudian dengan model SVM dilakukan proses klasifikasi dengan hasil kategori negatif 63,9%, netral 18%, dan positif 18,1%, dapat disimpulkan bahwa implementasi program ini telah menimbulkan adanya ketidakpuasan di kalangan masyarakat disaat munculnya perubahan peraturan. Proses pelabelan awal merupakan sebuah tantangan pada penilitian ini dimana memerlukan seorang ahli dalam menvalidasi setiap label komentar. Oleh karena itu, penelitian selanjutnya perlu lebih cermat dalam pelabelan awal untuk pembelajaran model dengan kecerdasan buatan yang lebih optimal, sehingga prosesnya menjadi lebih efisien dan otomatis tanpa memerlukan ahli untuk validasi. Selain itu, data yang digunakan masih berasal dari satu platform media sosial. Akan lebih baik jika penelitian selanjutnya melibatkan multiple platform untuk analisis yang lebih mendalam. Secara keseluruhan masyarakat memberikan respons yang lebih negatif terhadap perubahan Program TAPERA, evaluasi sentimen menunjukkan adanya kekhawatiran, terutama terkait dampaknya pada kondisi finansial pribadi dan transparansi pelaksanaan

program. Oleh karena itu, untuk meningkatkan efektivitas dan respons positif terhadap TAPERA, diperlukan perbaikan dalam sosialisasi program serta pengawasan yang lebih ketat, agar masyarakat tidak merasa terbebani dan memahami manfaat program ini dengan lebih jelas.

REFERENSI

[1] Y. M. De, "Analisis Kritis Program TAPERA 'Tabungan Perumahan Rakyat' Bagi Kehidupan Umat di Paroki Riam Batang Kalimantan Tengah," *J. Pendidik. Agama dan Teol.*, vol. 2, no. 3, pp. 57–73, 2024, doi: https://doi.org/10.59581/jpat-widyakarya.v2i3.3354.

[2] T. A. Nasution, "Analisis Yuridis Undang-Undang Tabungan Perumahan Rakyat Ditinjau Dari Perspektif Good Governance," *LEX Renaissan*, vol. 6, no. 4, pp. 833–846, 2021, doi: https://doi.org/10.20885/JLR.vol6.iss4.art13.

[3] M. Ihsan, A. Rofiq, and Khusnudin, "Polemik Tabungan Perumahan Rakyat (TAPERA): Sebuah kajian dengan pendekatan interdisipliner," *Gulawentah J. Stud. Sos.*, vol. 9, no. 1, pp. 72–78, 2024, doi: https://doi.org/10.25273/gulawentah.v9i1.20497.

[4] A. Perdana, A. Hermawan, and D. Avianto, "Analisis Sentimen Terhadap Isu Penundaan Pemilu di Twitter Menggunakan Naive Bayes Classifier," *J. SISFOKOM (Sistem Inf. dan Komputer)*, vol. 11, no. 2, pp. 195–200, 2022, doi: https://doi.org/10.32736/sisfokom.v11i2.1412.

[5] D. W. Ardras and A. Voutama, "Analisis Sentimen Anti Lgbt Di Indonesia Melalui Media Sosial Twitter," *J. Tek. (Jurnal Fak. Tek. Univ. Islam Lamongan)*, vol. 15, no. 1, pp. 23–28, 2023, doi: https://doi.org/10.30736/jt.v15i1.926.

[6] R. Gusti Wardhana, G. Wang, and F. Sibuea, "Penerapan Machine Learning Dalam Prediksi Tingkat Kasus Penyakit Di Indonesia," *J. Inf. Syst. Manag.*, vol. 5, no. 1, pp. 40–45, 2023, doi: https://doi.org/10.24076/joism.2023v5i1.1136.

[7] M. I. Alfandi, P. Adytia, and Wahyuni, "Analisis Sentimen Masyarakat Terhadap TAPERA Pada Media Sosial X Menggunakan Metode K-Nearest Neighbor," *SEBATIK*, vol. 28, no. 2, 2024, doi: 10.46984/sebatik.v28i2.0000.

[8] R. Hafil Muhammadi, T. Ginanjar Laksana, and A. Beladinna Arifa, "Combination of Support Vector Machine and LexiconBased Algorithm in Twitter Sentiment Analysis," *KHAZANAH Inform.*, vol. 8, no. 1, pp. 59–71, 2022, doi: https://doi.org/10.23917/khif.v8i1.15213.

[9] A. Maulidatur Rizqiyah and I. Kadek Dwi Nuryana, "Analisis Sentimen Masyarakat terhadap Kebijakan Iuran Tabungan Perumahan Rakyat (TAPERA) pada Platform X Menggunakan Algoritma Naïve Bayes Classifier dan Support Vector Machine," *J. Emerg. Inf. Syst. Bus. Intell.*, vol. 5, no. 3, pp. 298–306, 2024, [Online]. Available: https://ejournal.unesa.ac.id/index.php/JEISBI/article/view/64074

[10] A. Ananta Firdaus, A. Id Hadiana, and A. Kania Ningsih, "Klasifikasi Sentimen pada Aplikasi Shopee Menggunakan Fitur Bag of Word dan Algoritma Random Forest," *J. Multidiscip. Res. Dev.*, vol. 6, no. 5, pp. 1678–1683, 2024, doi: https://doi.org/10.38035/rrj.v6i5.994.

[11] M. utfi Pratama, Y. Vita Via, and E. Prakarsa Mandyartha, "Analisis Performansi Naive Bayes Dan Random Forest Terhadap Sentimen Kenaikan Harga Bbm Di Indonesia," *J. Teknol. Inf. dan Komun.*, vol. 18, no. 1, pp. 18–24, 2023, doi: https://doi.org/10.33005/scan.v18i1.3837.

[12] M. Iqbal, M. Afdal, and R. Novita, "Implementasi Algoritma Support Vector Machine untuk Analisa Sentimen Data Ulasan Aplikasi Pinjaman Online Di Google Play Store," *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 4, no. 4, pp. 1244–1252, 2024, doi: https://doi.org/10.57152/malcom.v4i4.1435.

[13] F. Adi Artanto, "Implementasi Algoritma Random Forest dan Model Bag of Words Dalam Analisis Sentimen Mengenai E-Materai," *J. Sains Teknol. dan Sist. Inf.*, vol. 4, no. 2, pp. 139–145, 2024, doi: https://doi.org/10.54259/satesi.v4i2.3240.

[14] N. Cholifah Sastya and I. Nugraha, "Penerapan Metode CRISP-DM dalam Menganalisis Data untuk Menentukan Customer Behavior di MeatSolution," *J. Pendidik. Dan Apl. Ind.*, vol. 10, no. 2, pp. 103–115, 2023, doi: https://doi.org/10.33592/unistek.v10i2.3079.

[15] M. Rafi Muttaqin, T. Iman Hermanto, and M. Agus Sunandar, "Penerapan K-Means Clustering Dan Cross-Industry Standard Process For Data Mining (Crisp-Dm) Untuk Mengelompokan Penjualan Kue," *KOMPUTASI J. Ilm. Ilmu Komput. dan Mat.*, vol. 19, no. 1, pp. 38–53, 2022, doi: 10.33751/komputasi.v19i1.3976.

[16] D. Pramudita, Y. Akbar, and T. Wahyudi, "Analisis Sentimen Terhadap Program Kartu Indonesia Pintar Kuliah Pada Media Sosial X Menggunakan Algoritma Naive Bayes," *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 4, no. 4, pp. 1420–1430, 2024, doi: https://doi.org/10.57152/malcom.v4i4.1565.

[17] A. Desiani *et al.*, "Penerapan Metode Support Vector Machine Dalam Klasifikasi Bunga Iris," *IJAI (Indonesian J. Appl. Informatics)*, vol. 7, no. 1, pp. 12–18, 2022, doi: https://doi.org/10.20961/ijai.v7i1.61486.

[18] A. Rahma Lestari, R. Santoso, and Suparti, "Analisis Sentimen Pengguna Online Travel Agent (Ota) Pada Perusahaan Pegipegi.Com Menggunakan Random Forest," *J. Gaussian*, vol. 12, no. 4, pp. 616–624, 2023, doi: https://doi.org/10.14710/j.gauss.12.4.616-624.

[19] A. Musthafa, D. Muriyatmoko, Taufiqurrahman, and S. Kamal Sholihin, "Deteksi Berita Salah Pada Pemilihan Umum Presiden 2024 Menggunakan Metode Naïve Bayes Berbasis Website," *J. Fasilkom*, vol. 14, no. 2, pp. 410–419, 2024, doi: https://doi.org/10.37859/jf.v14i2.7110.

[20] J. Josen A. Limbong, I. Sembiring, and K. Dwi Hartomo, "Analisis Klasifikasi Sentimen Ulasan Pada E-Commerce Shopee Berbasis Word Cloud Dengan Metode Naive Bayes Dan K-Nearest Neighbor," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 9, no. 2, pp. 347–355, 2022, doi: https://doi.org/10.25126/jtiik.2022924960.