BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 5, No 4, June 2025 | Hal 277-286
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v5i4.523

Model Machine Learning Untuk Analisis Sentimen Masyarakat Terhadap Kenaikan PPN di Media Sosial X

Ilham Ridho Pratama*, Yana Cahyana, Rahmat, Deden Wahiddin

Fakultas Ilmu Komputer, Teknik Informatika, Universitas Buana Perjuangan Karawang, Karawang, Indonesia
Email: 1*if21.ilhampratama@mhs.ubpkarawang.ac.id, 2yana.cahyana@ubpkarawang.ac.id, 3rahmat@ubpkarawang.ac.id, 4deden.wahiddin@ubpkarawang.ac.id
Email Penulis Korespondensi: if21.ilhampratama@mhs.ubpkarawang.ac.id

Abstrak−Studi ini mengkaji reaksi masyarakat terhadap rencana pemerintah Indonesia penyesuaian tarif PPN dari 11% menjadi 12% yang direncanakan mulai efektif pada tahun 2025. Kebijakan ini memicu beragam pendapat di kalangan warganet, khususnya layanan jejaring sosial X. Untuk menelusuri opini masyarakat, data dikumpulkan melalui teknik web *crawling* dalam kurun waktu Oktober hingga Desember 2024, menghasilkan 1.871 data. Kemudian dataset melalui praproses teks *cleaning*, *case folding*, *tokenize*, *stopword removal*, dan *stemming* dan dataset berkurang menjadi 1806. Selanjutnya, sebanyak 1000 data diberikan label, negatif, netral, positif secara manual oleh pakar bahasa guna memastikan setiap kalimat memiliki label yang sesuai. Data ini akan di gunakan untuk *test* dan *train*, kemudian sebanyak 806 data tanpa label digunakan sebagai pengujian akhir. Pada tahapan pembobotan kata dihitung melalui metode *Term Frequency-Inverse Document Frequency* (*TF-IDF*) digunakan untuk melakukan proses tersebut. Tiga algoritma *machine learning* diterapkan dalam penelitian ini untuk perbandingan performa klasifikasi, yaitu *Support Vector Machine* (*SVM*), *Random Forest*, dan *Decision Tree*. Berdasarkan hasil evaluasi, algoritma *SVM* mencatat tingkat akurasi tertinggi sebesar 94%, diikuti *Random Forest* dengan 93%, dan *Decision Tree* sebesar 91%. Temuan menunjukkan dominasi sentimen negatif, yang menandakan ketidakpuasan masyarakat terhadap kebijakan tersebut. Studi ini membuktikan bahwa teknik *machine learning* dapat digunakan secara efektif untuk menangkap persepsi publik melalui media sosial, yang pada akhirnya bisa menjadi tolak ukur bagi pemerintah untuk mengambil sebuah keputusan yang akan diberlakukan.

Kata Kunci: PPN 12%; Sentiment Analysis; Machine Learning; Social Media X; SVM; Random Forest; Decision Tree

Abstract−This study examines people's reactions to the Indonesian government's plan to adjust the VAT rate from 11% to 12%, which is scheduled to take effect in 2025. This policy triggered a variety of opinions among netizens, especially on the social networking service X. To explore public opinion, data was collected through web crawling techniques from October to December 2024, resulting in 1,871 records. Then the dataset was preprocessed by text cleaning, case folding, tokenization, stopword removal, and stemming, and the dataset was reduced to 1806. In addition, up to 1000 data will be manually labeled, negative, neutral, positive, by language experts to ensure that each sentence has the appropriate label. These data are used for testing and training, then up to 806 unlabeled data are used as final testing. At the word weighting stage, the Term Frequency-Inverse Document Frequency (TF-IDF) method is used to perform the process. In this study, three machine learning algorithms were used to compare the classification performance, namely Support Vector Machine (SVM), Random Forest, and Decision Tree. Based on the evaluation results, the SVM algorithm recorded the highest accuracy rate of 94%, followed by Random Forest with 93% and Decision Tree with 91%. The results showed a predominance of negative sentiments, indicating public dissatisfaction with the policy. This study proves that machine learning techniques can be effectively used to capture public perceptions through social media, which in turn can be a benchmark for the government to make decisions that will be enforced.

Keywords: PPN 12%; Sentiment Analysis; Machine Learning; Social Media X; SVM; Random Forest; Decision Tree

1. PENDAHULUAN

Di Indonesia, sistem perpajakan mencakup berbagai jenis pungutan pajak, di antaranya Pajak Pertambahan Nilai (PPN) yang dibebankan dalam setiap transaksi barang maupun jasa. PPN berperan sebagai salah satu sumber pendanaan utama bagi negara untuk membiayai proyek-proyek pembangunan serta meningkatkan pelayanan publik kepada masyarakat [1].

Pemerintah berencana menyesuaikan tarif PPN, yang semula 11%, menjadi 12%. Sejalan dengan reformasi perpajakan dalam UU No. 7/2021 terkait harmonisasi aturan pajak. Ketentuan ini dijelaskan secara rinci dalam Bab 4 Pasal 7 Ayat (2) UU HPP, di mana disebutkan bahwa kenaikan tersebut akan berlaku mulai 1 Januari 2025. Tujuan dari kebijakan ini guna meningkatkan penerimaan negara dalam mendukung proses stabilisasi ekonomi dan penguatan pembangunan negara [2]. Perubahan kebijakan ini mendapat *respons* yang beragam dari masyarakat, yang banyak diungkapkan melalui media sosial seperti X. Perkembangan media sosial telah mengubah cara individu menyampaikan pendapat, serta menjadi sarana yang memperkuat partisipasi masyarakat dengan memberikan akses informasi yang luas [3].

Survei *We Are Social* menunjukkan bahwa Indonesia menempati posisi keempat dunia dalam jumlah pengguna media sosial X pada Oktober 2023, dengan total pengguna sekitar 27,5 juta [4]. Dengan tingginya jumlah pengguna media sosial X di Indonesia, penyebaran informasi, termasuk mengenai isu kenaikan PPN, menjadi sangat cepat dan luas. Hal ini menyebabkan munculnya berbagai pendapat dari masyarakat mengenai kebijakan tersebut. Analisis sentimen menjadi penting untuk menggali persepsi masyarakat, baik yang pro maupun kontra, guna membantu pemerintah dalam mengevaluasi dampak kebijakan serta meningkatkan efektivitas komunikasi publik.

Untuk mengkaji tanggapan masyarakat atas rencana kenaikan PPN sebesar 12%, penelitian ini menerapkan pendekatan *machine learning* sebagai metode utama analisis. Teknologi ini memungkinkan pemrosesan data dalam skala besar yang diperoleh dari media sosial X, sehingga dapat mengidentifikasi kecenderungan opini publik terhadap kebijakan tersebut dengan menggunakan algoritma klasifikasi.

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 5, No 4, June 2025 | Hal 277-286
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v5i4.523

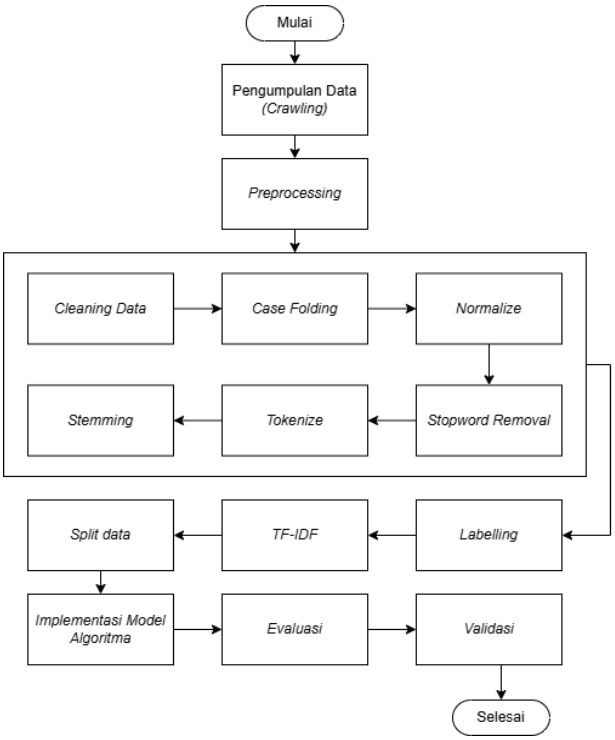
Penelitian oleh Jesica Kristovani Siagian dan Painem tentang sentimen kenaikan PPN di platform sosial X menggunakan model *Naïve Bayes* memberikan skor yang akurat dengan 74%, *precision* 83%, dan *recall* 5,2% dari 486 *tweet* yang dianalisis [5]. Studi lain algoritma *SVM* untuk menganalisis respon terhadap e-tilang mencatat akurasi 74,20%, namun masih lemah dalam mendeteksi sentimen netral [6]. Berliana Nur Isnayni dan tim pada tahun 2024 menggunakan metode *Random Forest* pada ulasan kafe di media daring dan mendapatkan akurasi sebesar 79% dengan presisi tinggi untuk sentimen negatif [7]. Penelitian lainnya oleh Puspita Dewi dkk. dengan algoritma *Decision Tree* terhadap ulasan aplikasi Canva menunjukkan akurasi 87% [8]. sementara Muhayat dkk. (2023) yang menerapkan *SVM* pada komentar youtube dengan *TF-IDF* memperoleh akurasi 86% dan *precision* 87% [9]. Studi-studi ini mengindikasikan bahwa *SVM*, *Random Forest*, dan *Decision Tree* memiliki kapabilitas yang baik dalam analisis sentimen, meski tiap metode tetap memiliki kekurangan.

Penelitian ini secara khusus lebih berfokus kepada pendapat masyarakat terhadap kenaikan PPN pada jejaring social X, menawarkan pendekatan komparatif menggunakan ketiga algoritma tersebut secara bersamaan (*SVM*, *Random Forest*, dan *Decision Tree*) dalam menganalisis respons masyarakat terhadap kebijakan kenaikan PPN sebesar 12%. Berdasarkan penelitian terdahulu, algoritma *SVM*, *Random Forest*, dan *Decision Tree* dipilih karena menunjukkan kinerja yang kompetitif dalam tugas analisis sentimen, baik dari segi akurasi maupun presisi. dengan jumlah data sebesar 1.806 *tweet* dari media sosial X.

Tujuan penelitian ini guna mengevaluasi respon masyarakat pada rencana peningkatan tarif PPN melalui percakapan di media sosial X. Dengan memahami pola sentimen masyarakat secara lebih akurat, penelitian ini dapat berkontribusi dalam mendukung proses pengambilan kebijakan publik yang lebih responsif dan berbasis data. Dampak utama yang diharapkan adalah terciptanya kebijakan ekonomi yang tidak hanya efisien, tetapi juga mempertimbangkan aspirasi dan kekhawatiran masyarakat. Selain itu, hasil penelitian ini dapat menjadi referensi penting bagi pembuat kebijakan dalam menyusun strategi komunikasi yang lebih efektif untuk mengurangi resistensi publik terhadap kebijakan yang kontroversial, sekaligus meningkatkan kepercayaan masyarakat terhadap pemerintah. Pada tingkat yang lebih luas, penelitian ini juga memberikan kontribusi akademik dengan menunjukkan bagaimana machine learning dapat digunakan untuk menganalisis isu-isu sosial dan ekonomi, sehingga menjadi pendekatan yang relevan untuk studi serupa di masa depan.

2. METODOLOGI PENELITIAN

Berikut Gambar 1 merupakan tahapan dari penelitian yang dilakukan.



Gambar 1. Alur keseluruhan penelitian

Gambar 1 menunjukan alur penelitian, dimulai dengan tahap pengumpulan data menggunakan teknik *crawling*, yaitu mengumpulkan informasi dari berbagai sumber, seperti file, *database*, atau API, dengan tujuan untuk memungkinkan analisis yang diperlukan untuk pengembangan dan penelitian [10].

Data yang digunakan diperoleh dengan memanfaatkan platform *Google Colab* dan menggunakan pustaka *Python* bernama *tweet-harvest*. Data yang digunakan pada penelitian ini sebanyak 1806, dengan menggunakan proses

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 5, No 4, June 2025 | Hal 277-286
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v5i4.523

pengambilan data dilakukan di media sosial X selama periode 29 Oktober hingga 17 Desember 2024 dengan kata kunci pencarian "PPN 12%". Hasil *crawling* disimpan dalam format .csv. Media sosial X, yang kini sangat digemari oleh kalangan mahasiswa, memungkinkan pengguna terhubung dengan berbagai informasi secara global [11].

Data yang diperoleh kemudian melalui proses *preprocessing* untuk meningkatkan kualitas data sebelum dianalisis. Proses *preprocessing* mencakup beberapa langkah, yaitu :

a. *Cleaning*
   Membersihkan data yang kurang relevan seperti url, *hashtag*, dan *mention*[8].
b. *Case folding*
   Merubah seluruh struktur kalimat menjadi huruf kecil, guna mempermudah proses pembuatan model [12].
c. *Normalizing*
   Merubah kata yang tidak baku menjadi baku, seperti "ngga" diubah menjadi "tidak" dan kata yang di singkat seperti "tdk" menjadi "tidak" [8].
d. *Stopword removal*
   Penghapusan kata-kata yang sering muncul tetapi tidak relevan seperti "tidak", "dan". Menggunakan *libarary* sastrawi yang terdapat di dalam *google colab* [13].
e. *tokenizing*
   Tokenizing dilakukan setelah tahap *cleaning*, *normalizing*, dan *stopword* removal. Tahap ini memecah kalimat ke dalam kata-kata individu untuk memudahkan pemrosesan model dan analisis, *tokenizing* dilakukan menggunakan metode pemisahan sederhana berbasis spasi pada *Python*, tanpa menggunakan pustaka *tokenizer* khusus [14].
f. *Stemming*
   Proses *stemming* dilakukan menggunakan pustaka sastrawi, sebuah proses mencari dasar dari setiap kata sehingga terbentuk menjadi kata dasarnya dengan cara menghapus imbuhan seperti awalan, akhiran, dan sisipan untuk mendapatkan bentuk dasar kata [15].

Setelah *preprocessing* selesai, dari total 1806 data, dibagi menjadi 3 : *train*, *test*, dan validasi.

Tabel 1. Pembagian dataset

Dataset	Jumlah
Train dan *test*	1000
Validasi	806
Jumlah	1806

Data *train* dan *test* melalui tahap *labelling*, dan dibagi menjadi tiga sentimen: negatif, netral, positif. Proses pelabelan dataset dilakukan oleh pakar bahasa berdasarkan pemahaman linguistik dan konteks setiap sampel. Kriteria pelabelan mengacu pada kesesuaian makna dan penggunaan bahasa dalam teks. Seluruh proses dilakukan oleh pakar yang sama untuk menjaga konsistensi antar label, sehingga meningkatkan akurasi dan keandalan model *machine learning* (ML). ML merupakan cabang ilmu komputer yang berorientasi pada pengembangan algoritma guna membantu komputer memahami dan belajar dari data, atau yang sering disebut sebagai belajar dari data. Dalam hal ini, *machine learning* digunakan untuk membangun model pembelajaran yang dapat mengekstrak informasi dari kumpulan data yang besar, dengan cara yang paling efektif. Esensinya, *machine learning* bertujuan untuk membangun model yang dapat mencerminkan pola dalam data tanpa mengurangi makna yang terkandung di dalamnya [16].

Sebanyak 1000 dataset *train* dan *test* yang sudah diberikan label akan di bagi ke dalam fraksi 80:20, dimana dari 1000 data 80% akan digunakan sebagai *train* dan 20% sebagai *test*. Sementara dataset validasi berujumlah 806, terdiri dari data yang belum diberi label.

Selanjutnya adalah transformasi data yang akan menggunakan *TF-IDF* untuk pembobotan kata, Nilai diberikan kepada setiap kata yang telah melewati tahap perancangan sebelumnya dalam proses pembobotan kata. Persamaan (1) dan (2) menunjukkan metode *TF-IDF* :

$$\text{IDF(w)} = \log\left(\frac{N}{\text{DF(w)}}\right), \quad (1)$$

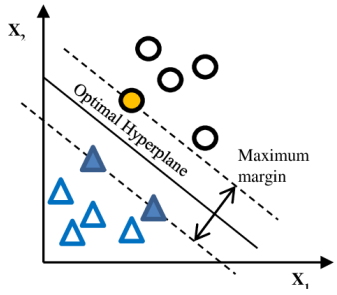
$$W_{ij} = TF_{ij} \text{x } \log\left({}^{D}/_{DFj}\right). \quad (2)$$

Langkah-langkah dalam proses ini meliputi perhitungan frekuensi kemunculan kata (TF), penjumlahan kemunculan kata dalam seluruh dokumen (DF), perhitungan IDF, dan akhirnya menentukan bobot dokumen berdasarkan semua kata [17].

Selanjutnya, model *machine learning* dikembangkan menggunakan tiga algoritma berikut :

a. *Support Vector Machine* (*SVM*)
   SVM menggunakan pendekatan mencari garis pembatas optimal (*hyperplane*) mencakup dua kelas data. Dalam konteks analisis sentimen, *SVM* membantu membedakan positif dan negatif dengan cara memaksimalkan jarak antar titik data terhadap garis pemisah. Model ini efektif untuk klasifikasi karena mampu menangani data linier maupun non-linier [18].

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 5, No 4, June 2025 | Hal 277-286
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v5i4.523

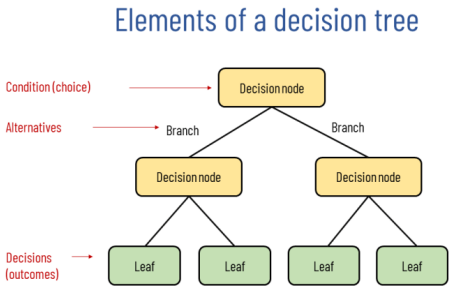


Gambar 2. Margin *hyperplane*

Teknik prediksi untuk klasifikasi dan regresi dikenal sebagai *SVM* [6]. Pada tugas klasifikasi sentimen terhadap data tweet, *SVM* dapat disesuaikan dengan memilih kernel yang sesuai berdasarkan distribusi data. Kernel yang digunakan pada penelitian ini yaitu linear, karena data yang digunakan dianggap dapat dipisahkan secara linier di ruang fitur. Kernel linear memiliki keunggulan dalam kecepatan komputasi dibandingkan kernel lainnya, terutama ketika jumlah fitur yang diekstraksi dari data teks sangat besar, seperti pada kasus analisis sentimen terhadap tweet.

b. *Decision Tree* (DT)

Decision tree adalah model prediktif berbasis struktur pohon. Data akan dibagi berdasarkan pengujian terhadap fitur tertentu dan mengarahkannya melalui cabang-cabang pohon hingga mencapai simpul daun yang menentukan hasil akhir. Pohon dimulai dengan simpul akar.



Gambar 3. Cara kerja *Decision Tree*

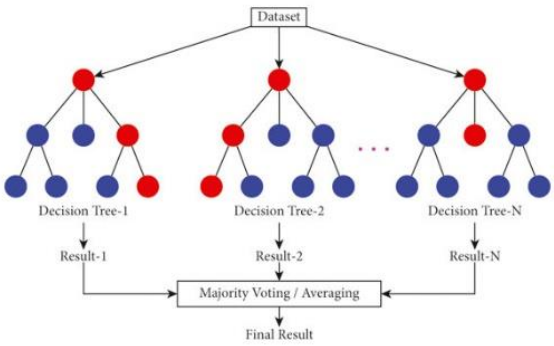
Cara kerja *Decision Tree* dimulai dari node akar, yang membagi data berdasarkan fitur dengan informasi terbaik, setiap hasil pengujian menentukan cabang yang akan dipilih berikutnya, proses ini berulang hingga semua data dalam satu cabang termasuk dalam satu kelas atau mencapai kondisi penghentian, simpul daun mewakili hasil akhir klasifikasi atau prediksi. Model ini banyak digunakan karena dapat menangani berbagai jenis data dan menghasilkan hasil yang mudah dipahami [19].

$$\text{Gini impurity} = 1 - \sum_{i=1}^{c} p_i^2 \quad (3)$$

Gini impurity digunakan pada studi ini karena mampu menghitung ketidakmurnian suatu simpul dengan efisien. Kriteria ini mengukur probabilitas suatu data salah klasifikasi jika dipilih secara acak, sehingga membantu dalam membagi data dengan akurasi tinggi.

c. *Random Forest* (RF)

Metode *Random Forest* merupakan algoritma pembelajaran ensemble yang memanfaatkan struktur pohon dalam prosesnya. Dalam penerapannya, pohon keputusan dibangun dengan memisahkan data secara acak. Pemilihan sampel data dilakukan menggunakan mekanisme voting berdasarkan output dari masing-masing *Decision Tree*.



Gambar 4. Cara kerja algortima *Random Forest*

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 5, No 4, June 2025 | Hal 277-286
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v5i4.523

Teknik *bagging* dengan atribut yang dipilih secara acak dapat digunakan untuk membangun metode *Random Forest*. Untuk membuat pohon keputusan, algoritma klasifikasi dan regresi pohon (CART) dapat digunakan. Dalam metode ini, pohon-pohon tersebut akan tumbuh tanpa pemangkasan hingga mencapai tingkat maksimum mereka, menghasilkan kumpulan pohon yang disebut hutan/*forest* [7].
Pada studi ini, jumlah pohon dalam *Random Forest*, yang dikenal sebagai *n_estimators*, diatur menjadi 100, yang merupakan nilai *default* dalam pustaka *Scikit-learn*. Jumlah ini dipilih karena memberikan keseimbangan antara akurasi dan efisiensi komputasi. Dengan 100 pohon, algoritma mampu menghasilkan prediksi yang stabil tanpa memerlukan waktu komputasi yang terlalu lama.

Untuk mengukur kinerja ketiga model maka digunakan confusion matrix, sebagai tabel perbandingan antara prediksi model dengan label sebenarnya. Model yang sudah dibuat akan dievaluasi menggunakannya, model yang baik akan memiliki nilai tinggi pada diagonal utama (*True Positives* dan *True Negatives*) serta nilai rendah atau mendekati nol di luar diagonal. Berikut adalah representasi confusion matrix untuk kasus 2x2:
Dari *confusion matrix*, terdapat tiga metrik evaluasi utama yang sering digunakan, yaitu:

a. Akurasi = (TP + TN) / (TP + TN + FP + FN) (4)
b. Presisi = TP / (TP + FP) (5)
c. *Recall* = TP / (TP + FN) (6)

Hasil akurasi, *precision*, dan *recall* umumnya dinyatakan dalam bentuk persentase dengan rentang 0 hingga 100%. Suatu sistem dikatakan memiliki kinerja yang baik jika ketiga metrik tersebut memiliki nilai yang tinggi [20]. Model yang sudah selesai di buat akan digunakan untuk pengujian. Validasi akhir dilakukan menggunakan 806 data baru yang tidak dipakai sebelumnya, untuk mengukur generalisasi model pada data yang belum pernah dilatih.

3. HASIL DAN PEMBAHASAN

3.1 Pengumpulan data

Studi mengumpulkan data tweet dati X memanfaatkan kata kunci “ppn 12%” berbahasa Indonesia. Data yang dikumpulkan berasal dari pengguna X dari 29 Oktober 2024 hingga 17 Desember 2024.

	conversation_id_str	created_at	favorite_count	full_text	id_str	image_url	in_reply_to_screen_name	lang	location	quote_count	reply_count	retweet_count
0	1.868108e+18	Sun Dec 15 14:48:26 +0000 2024	0	@txtdrimedia Ppn 12% KOCAK	1.868307e+18	NaN	txtdrimedia	in	nomin nation.	0	0	0
1	1.868108e+18	Sun Dec 15 14:45:01 +0000 2024	0	@txtdrimedia pajak motor nambah 2 ppn naek 12%...	1.868306e+18	NaN	txtdrimedia	in	Jakarta Selatan, DKI Jakarta	0	0	0
2	1.868306e+18	Sun Dec 15 14:42:20 +0000 2024	0	OJK: PPN 12% Berdampak Temporer https://t.co/K...	1.868306e+18	NaN	NaN	in	Jakarta, Indonesia	0	0	0
3	1.868304e+18	Sun Dec 15 14:41:34 +0000 2024	0	@aHeartShaker oh iyaya ppn naik jadi 12% . .	1.868305e+18	NaN	aHeartShaker	in	LAYOUT!	0	1	0
4	1.868305e+18	Sun Dec 15 14:40:26 +0000 2024	35	Colek Prabowo soal Wacana Kenaikan PPN 12 Pers...	1.868305e+18	NaN	NaN	in	NaN	0	0	21
...	...	...	...	...	...	...	...	...	...	...	...	...
1866	1.869060e+18	Tue Dec 17 16:31:33 +0000 2024	0	Menteri Keuangan Sri Mulyani mengumumkan pembe...	1.869060e+18	NaN	NaN	in	NaN	0	0	0

Gambar 5. Crawling data

Data yang telah diambil sebanyak 1871. Data yang tersedia masih tidak terstruktur dan penuh dengan tanda baca, angka, simbol, dan kata-kata tidak biasa. Oleh karena itu, diperlukan tahap *preprocessing* untuk membersihkan data ini sebelum melanjutkan ke proses klasifikasi.

3.2 *Data Selection*

Data yang dikumpulkan disimpan dalam file berformat .csv dan berisi beberapa kolom. Namun, untuk keperluan analisis, hanya kolom berisi teks lengkap (*full_text*) yang dipilih sebagai fokus utama :

```
Index(['conversation_id_str', 'created_at', 'favorite_count', 'full_text',
       'id_str', 'image_url', 'in_reply_to_screen_name', 'lang', 'location',
       'quote_count', 'reply_count', 'retweet_count', 'tweet_url',
       'user_id_str', 'username'],
      dtype='object')
```

Gambar 6. Kolom sebelum dilakukan data *selection*

Untuk memudahkan analisis data, hanya kolom full_text yang akan diambil.

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 5, No 4, June 2025 | Hal 277-286
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v5i4.523

3.3 *Pre-processing Data*

Tahapan *preprocessing* data merupakan tahapan yang cukup penting untuk membersikan data yang tidak diperlukan, hingga analisis sentiment yang dilakukan dapat lebih akurat dan efektif. Tahapan pertama yang dilakukan yaitu *cleaning*, yakni menghapus tanda baca yakni menghapus tanda baca, mention, hashtag, URL, dan elemen tidak relevan lainnya. Proses ini juga mencakup pengecekan apakah terdapat duplikasi atau data kosong.

Tabel 2. Data *cleaning*

Data kotor	Data bersih
@txtdrimedia Ppn 12% KOCAK	ppn 12 persen kocak
@aHeartShaker oh iyaya ppn naik jadi 12%	oh iyaya ppn naik jadi 12%
Colek Prabowo soal Wacana Kenaikan PPN 12 Persen	colek prabowo soal wacana kenaikan ppn 12 persen

Setelah proses *cleaning* data, proses *case folding* diterapkan untuk mengonversi seluruh huruf. Selanjutnya, proses *normalizing* guna memperbaiki kata singkatan. diikuti oleh penghapusan *stopword* seperti "dan", "yang", dan sebagainya. Selanjutnya, tahap *tokenizing* dilakukan dengan memisahkan kalimat menjadi unit-unit kata. Tahap akhir, *stemming* berfungsi untuk mereduksi kata ke bentuk dasarnya, seperti mengubah “membaca” menjadi “baca”, guna menyederhanakan variasi kata dan meningkatkan akurasi dalam analisis. Setelah *preprocessing* selesai, jumlah data berkurang dari 1871 menjadi 1806.

3.4 Pelabelan Data

Labelling dataset dilakukan oleh pakar bahasa untuk memastikan setiap sampel memiliki label yang sesuai, sehingga meningkatkan akurasi dan keandalan model *machine learning*. Dataset diklasifikasikan menjadi positif, netral, dan negatif. Sebanyak 1.000 data dilabeli oleh pakar, sementara 806 data digunakan untuk pengujian akhir.

Tabel 3. Hasil pelabelan

Kelas/label	Jumlah
Positif	136
Netral	25
Negatif	839
Total	1.000

3.5 *Data Transformation*

Setelah pelabelan, data ditransformasi menggunakan *TF-IDF* untuk pembobotan kata. Fungsi *TfidfVectorizer* dari *sklearn* digunakan untuk menghasilkan daftar istilah yang diatur alfabetis. Bobot ditentukan berdasarkan kosakata yang paling sering ditemukan dalam dokumen, yang menunjukkan tingkat relevansi kata tersebut terhadap konteks dokumen.

3.6 Klasifikasi

Ada tiga algoritma yang digunakan yaitu *SVM*, *Random Forest*, *Decision Tree*. Dataset sebanyak 1.000 entri dibagi dalam rasio 80:20:

Tabel 4. Pembagian dataset

Persentase pada data		Jumlah	
Latih	Uji	Latih	Uji
80%	20%	800	200
Total		1.000	

Pada tabel 4 merupakan persentase data dan jumlah data untuk pembuatan model.

3.7 Evaluasi

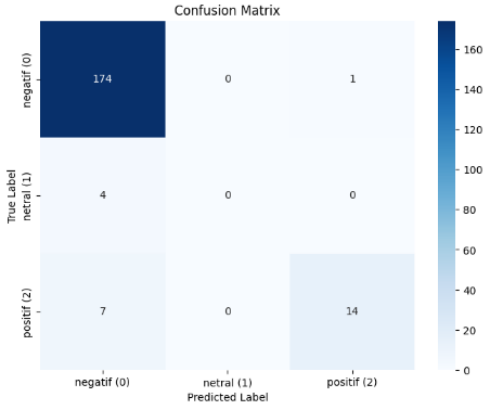
Setelah model berhasil dibangun, hasil evaluasi dari ketiga algoritma yang digunakan dalam skenario pembagian data 80:20 sebagai berikut :

Tabel 5. Hasil evaluasi model

Model	Akurasi	Presisi	Recall	F1-score
SVM	0,94	0,92	0,94	0,93
Random Forest	0,93	0,93	0,93	0,92
Decision Tree	0,91	0,91	0,91	0,92

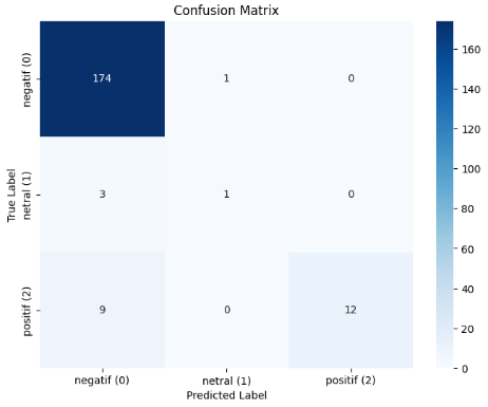
BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 5, No 4, June 2025 | Hal 277-286
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v5i4.523

Dari Tabel 5, hampir seluruh model yang digunakan menunjukan nilai yang cukup baik. Nilai terendah ditemukan pada model dengan menggunakan algortima *Decision Tree* dengan akurasi 91%, sedangkan nilai tertinggi terdapat pada model algortima *SVM* dengan akurasi sebesar 94%.



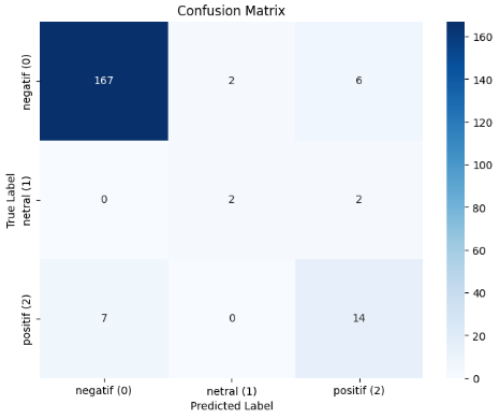
Gambar 7. *SVM*

Gambar 7 menunjukkan *confusion matrix* di mana model *SVM* sangat akurat untuk kelas negatif (174 benar, 1 salah) tetapi gagal memprediksi kelas netral (semua salah sebagai negatif). Untuk kelas positif, 14 prediksi benar dan 7 salah sebagai negatif, menunjukkan model lebih unggul dalam mengenali kelas negatif dibandingkan netral dan positif.



Gambar 8. *Random forest*

Gambar 8 menunjukan model *Random Forest*, di mana model menunjukkan akurasi tinggi pada kelas negatif (174 prediksi benar dan 1 salah). Untuk kelas netral, hanya terdapat 1 prediksi yang benar, sedangkan pada kelas positif, terdapat 12 prediksi benar dengan 9 prediksi salah yang diklasifikasikan sebagai negatif. Model ini menunjukkan performa yang lebih baik pada kelas negatif dibandingkan dengan kelas netral dan positif.



Gambar 9. *Decision tree*

Gambar 9 memperlihatkan evaluasi model *Decision Tree*. Kelas negatif, terdapat 167 prediksi benar, 2 salah sebagai netral, dan 6 sebagai positif. Pada kelas netral, hanya 2 prediksi benar, sementara pada kelas positif terdapat 14

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 5, No 4, June 2025 | Hal 277-286
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v5i4.523

benar dan 7 salah sebagai negatif. Model unggul pada kelas negatif tetapi memiliki banyak kesalahan pada kelas netral dan positif.

3.8 Pengujian

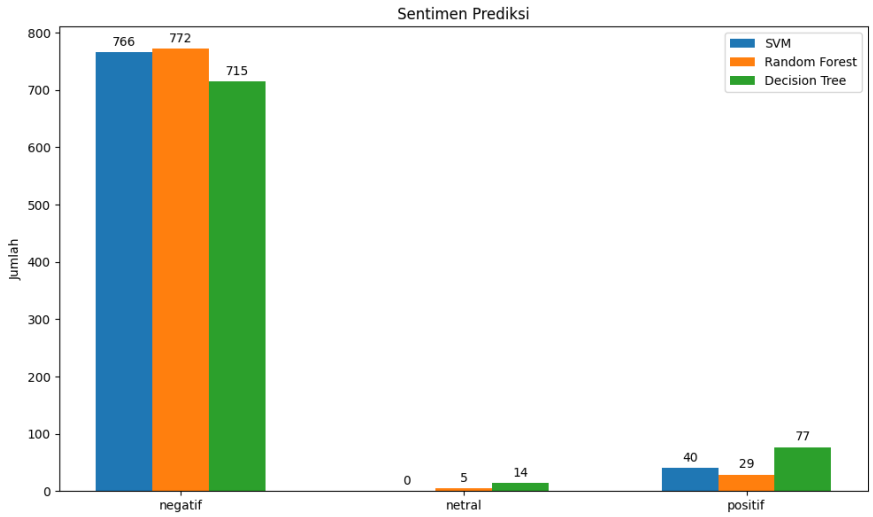
Setelah selesai membuat model, pada tahap ini sebanyak 806 data belum dilengkapi dengan label akan di gunakan untuk pengujian menggunakan model yang sudah di buat.

4	kenaikan ppn 12 persen mendukung program pemer...	positif	positif	positif
...	...	...	...	...
801	bulan november kemaren beras premium bahien 25...	negatif	negatif	negatif
802	lha terus sabun mandi deterjen baju daster pak...	negatif	negatif	negatif
803	tetep aja 300rb ga kecil lu emang tau struggle...	negatif	negatif	negatif
804	as kpopers kerja jadi ibu rumahtangga mau turu...	negatif	negatif	positif
805	ppn 12 persen memberikan manfaat nyata untuk m...	negatif	negatif	negatif

806 rows × 4 columns

Gambar 10. Hasil pelabelan menggunakan model

Model yang telah dibangun menunjukkan kinerja yang baik dalam melakukan pelabelan terhadap data baru yang memiliki topik serupa dengan data pelatihan. Hal ini terlihat dari hasil pelabelan yang konsisten dan relevan, yang mengindikasikan bahwa model mampu mengenali pola-pola yang terdapat pada data baru meskipun sebelumnya belum pernah melihat data tersebut.



Gambar 11. Distribusi sentiment SVM, RF, DT

Pada Gambar 11 menunjukan hasil distribusi dari pelabelan menggunakan model yang sudah di buat, diperoleh bahwa sentimen negatif mendominasi pada ketiga algoritma yang digunakan, yaitu *SVM*, *Random Forest*, dan *Decision Tree*. Algoritma *Random Forest* mencatat jumlah prediksi sentimen negatif tertinggi sebanyak 772, diikuti oleh *SVM* sebanyak 766, dan *Decision Tree* sebanyak 715. Sementara itu, sentimen netral terdeteksi dalam jumlah yang sangat kecil, dengan *Decision Tree* memprediksi 14 data, *Random Forest* sebanyak 5 data, dan *SVM* tidak memprediksi data apapun sebagai netral. Untuk sentimen positif, *Decision Tree* mencatat prediksi terbanyak sebanyak 77 data, diikuti oleh *SVM* sebanyak 40, dan *Random Forest* sebanyak 29. Hasil ini menunjukkan bahwa secara umum, respons masyarakat terhadap kebijakan kenaikan PPN cenderung bernada negatif, dengan proporsi sentimen netral dan positif yang relatif rendah.

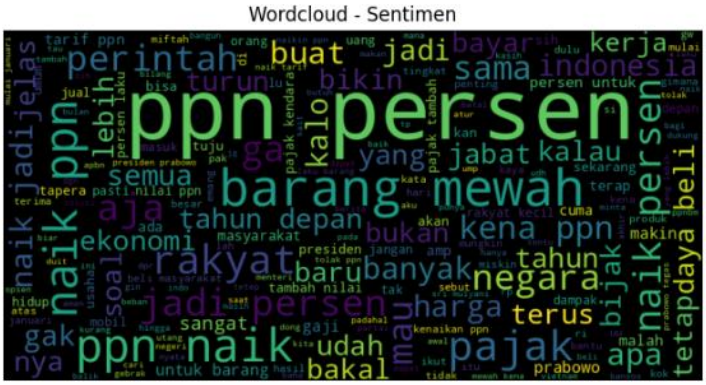
3.9 Pembahasan

Hasil penelitian ini menunjukkan bahwa algoritma *Support Vector Machine* (*SVM*) memberikan performa terbaik dalam klasifikasi sentimen terhadap data ulasan, dengan akurasi yang lebih tinggi dibandingkan algoritma *Decision Tree* dan *Random Forest*. Distribusi sentimen dalam penelitian ini menunjukkan kecenderungan dominasi sentimen negatif.

Pada penelitian Kristovani Siagian dan Painem dengan topik yang sama yaitu kenaikan PPN menggunakan algoritma *Naïve Bayes* diperoleh skor akurat dengan 74%, *precision* 83%, dan *recall* 5,2% dari 486 *tweet* yang dianalisis. Sementara pada penelitian ini menggunakan algortima *SVM* yang diperoleh akurasi 94%, diikuti oleh *Random Forest* (93%) dan *Decision Tree* (91%) dari 1000 *tweet* yang di analisis.

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 5, No 4, June 2025 | Hal 277-286
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v5i4.523

Kinerja *SVM* dalam penelitian ini dapat dijelaskan melalui karakteristik algoritma yang mampu menangani data berdimensi tinggi secara efisien, yang menjadi ciri khas data teks. *SVM* menggunakan *hyperplane* optimal untuk memisahkan kelas. Sementara itu, performa *Random Forest* sedikit lebih rendah karena meskipun algoritma ini mampu menangani data besar dan melakukan *ensembel* dari banyak pohon keputusan, namun sensitivitasnya terhadap data yang tidak seimbang dapat memengaruhi hasil klasifikasi. Sedangkan *Decision Tree* menunjukkan performa yang lebih rendah karena cenderung mengalami *overfitting*, terutama pada data teks dengan banyak fitur, sehingga mengurangi akurasi pada data uji.



Gambar 12. *Wordcloud keyword*

Pada Gambar 12 *Wordcloud* menunjukkan kata-kata yang sering ditemukan dalam data sentimen terhadap kebijakan kenaikan PPN 12%. Kata-kata seperti "ppn", "persen", "barang", "mewah", "naik", dan "negara" muncul dengan ukuran yang lebih besar, menunjukkan bahwa kata-kata tersebut memiliki frekuensi tinggi dalam teks. Ini mencerminkan fokus masyarakat pada dampak kenaikan PPN terhadap harga barang, ekonomi, dan kehidupan sehari-hari. Kata-kata seperti "rakyat" dan "pemerintah" menunjukkan adanya perhatian terhadap hubungan antara kebijakan dan masyarakat.

4. KESIMPULAN

Penelitian ini bertujuan untuk mengevaluasi respons masyarakat terhadap rencana kenaikan tarif PPN melalui percakapan di media sosial X. Hasil analisis sentimen menunjukkan bahwa mayoritas respons adalah negatif, dengan algoritma *SVM* menunjukkan kemampuan terbaik dalam mengklasifikasikan sentimen tersebut. *SVM* menghasilkan akurasi sebesar 94%, sementara *Random Forest* dan *Decision Tree* masing-masing mencatatkan akurasi 93% dan 91%. Dari analisis yang dilakukan, mayoritas sentimen yang terdeteksi bersifat negatif, mencerminkan rasa tidak setuju publik terhadap kebijakan tersebut. Dominasi sentimen negatif ini mengindikasikan adanya penolakan atau kekhawatiran yang signifikan di kalangan pengguna media sosial X terhadap kebijakan kenaikan PPN, yang perlu dipertimbangkan oleh pemerintah dalam strategi komunikasi dan implementasi kebijakan. Studi ini menunjukkan bahwa teknik *machine learning* dapat diterapkan secara efektif untuk menganalisis persepsi publik di media sosial, yang pada akhirnya dapat menjadi acuan bagi pemerintah dalam mengambil keputusan yang akan diterapkan. Namun, penelitian ini juga mengungkap beberapa keterbatasan. Pertama, klasifikasi sentimen netral dan positif masih kurang optimal, yang dapat disebabkan oleh ketidakseimbangan data dalam kategori tersebut, karena model cenderung lebih terlatih untuk mengenali pola pada kelas yang jumlah datanya dominan, dalam hal ini sentimen negatif. Akibatnya, model memiliki bias terhadap kelas tersebut dan kurang mampu mengenali ciri khas dari kelas netral dan positif yang jumlah datanya lebih sedikit, sehingga performa klasifikasi pada kedua kelas tersebut menjadi rendah. Kedua, penelitian ini hanya menggunakan data dari media sosial X selama periode terbatas, sehingga hasilnya mungkin tidak sepenuhnya merepresentasikan opini publik secara keseluruhan. Ketiga, metode *preprocessing* yang digunakan meskipun efektif, masih bisa ditingkatkan untuk meminimalkan kesalahan dalam analisis sentimen. Untuk penelitian selanjutnya, disarankan untuk meningkatkan jumlah dan variasi dataset, khususnya untuk kategori netral dan positif, guna mengurangi bias dalam model. Selain itu, eksplorasi algoritma lain, seperti *Gradient Boosting* atau *Neural Networks*, dapat dilakukan untuk membandingkan performa dan meningkatkan akurasi klasifikasi. Penambahan fitur seperti analisis temporal untuk melihat tren sentimen seiring waktu juga dapat menjadi nilai tambah dalam memahami perubahan opini publik. Dengan perbaikan ini, diharapkan model dapat memberikan hasil yang lebih representatif dan akurat untuk mendukung pengambilan keputusan berbasis data.

REFERENCES

[1] N. Fauziah, "Analisis Sentimen Publik Terhadap Kenaikan Tarif PPN di Indonesia dengan Pendekatan VADER," *Jurnal Akuntansi dan Keuangan*, vol. 12, no. 2, p. 228, Sep. 2024, doi: 10.29103/jak.v12i2.16796.

[2] I. M. Putri, "Kenaikan Ppn 12% Dan Dampaknya Terhadap Eknomi," *Jurnal Ilmiah Manajemen, Ekonomi, & Akuntansi (MEA)*, vol. 8, no. 2, pp. 934–944, 2024, doi: https://doi.org/10.31955/mea.v8i2.4077.

[3] S. Styawati and F. Ariany, "Sistem Monitoring Tumbuh Kembang Balita/Batita di Tengah Covid-19 Berbasis Mobile," *Jurnal Informatika Universitas Pamulang*, vol. 5, no. 4, p. 490, Dec. 2021, doi: 10.32493/informatika.v5i4.7067.

[4] L. Nursinggah, T. Mufizar, and U. Perjuangan, "Analisis Sentimen Pengguna Aplikasi X Terhadap Program Makan Siang Gratis Dengan Metode Naïve Bayes Classifier," *J. Inform. dan Tek. Elektro Ter*, vol. 12, no. 3, 2024, doi: http://dx.doi.org/10.23960/jitet.v12i3.4336.

[5] J. Siagian and P. Painem, "Analisis Sentimen Masyarakat Indonesia Terhadap Rencana Kenaikan Ppn Menjadi 12% Di Media Sosial Twiter/X Menggunakan Metode Naïve Bayes," in *Prosiding Seminar Nasional Mahasiswa Fakultas Teknologi Informasi (SENAFTI)*, 2024, pp. 779–786. Accessed: May 04, 2025. [Online]. Available: https://senafti.budiluhur.ac.id/senafti/article/view/1499

[6] D. Oktavia, Y. R. Ramadahan, and M. Minarto, "Analisis Sentimen Terhadap Penerapan Sistem E-Tilang Pada Media Sosial Twitter Menggunakan Algoritma Support Vector Machine (SVM)," *KLIK: Kajian Ilmiah Informatika dan Komputer*, vol. 4, no. 1, pp. 407–417, 2023, Accessed: May 04, 2025. [Online]. Available: https://djournals.com/klik/article/view/1040

[7] Berliana Nur Isnayni, Nurirwan Saputra, and Tri Hastono, "SENTIMENT ANALYSIS OF COFFEE SHOP REVIEWS USING RANDOM FOREST CLASSIFIER METHOD," *JTH: Journal of Technology and Health*, vol. 1, no. 4, pp. 233–244, May 2024, doi: 10.61677/jth.v2i2.152.

[8] Wulandari Wulandari, Nofiyani Nofiyani, and Yesi Puspita Dewi, "Analisis Sentimen terhadap Ulasan Aplikasi Canva di Play Store dengan Menggunakan Pendekatan Lexicon dan Algoritma Decision Tree," *Jurnal Ticom: Technology of Information and Communication*, vol. 13, no. 2, pp. 57–63, Jan. 2025, doi: 10.70309/ticom.v13i2.133.

[9] T. Muhayat, A. Fauzi, and J. Indra, "Analisis Sentimen Terhadap Komentar Video Youtube Menggunakan Support Vector Machines," *Progresif: Jurnal Ilmiah Komputer*, vol. 19, no. 1, p. 231, Feb. 2023, doi: 10.35889/progresif.v19i1.1060.

[10] N. A. R. Putri, "Analisis Jaringan pada Media Sosial X dengan# Boikot Menggunakan Social Network Analysis," *IJITECH: Indonesian Journal of Information Technology*, vol. 2, no. 1, pp. 11–15, 2024, Accessed: May 04, 2025. [Online]. Available: https://ojisnu.isnuponorogo.org/index.php/ijitech/article/view/79

[11] Rahmania Mustaqlillah, Okky Widyaningtyas, and Tri Wantoro, "Efektivitas Penggunaan Twitter Sebagai Sarana Peningkatan Berpikir Kritis Mahasiswa Ilmu Komunikasi," *MUKASI: Jurnal Ilmu Komunikasi*, vol. 2, no. 1, pp. 18–28, Feb. 2023, doi: 10.54259/mukasi.v2i1.1346.

[12] M. R. Pratama, A. Fauzi, D. Wahiddin, and A. R. Pratama, "Analisis Sentimen Kebijakan Pembelian Gas 3 Kg dengan KTP Menggunakan Naïve Bayes," *Jutisi : Jurnal Ilmiah Teknik Informatika dan Sistem Informasi*, vol. 13, no. 2, p. 1338, Aug. 2024, doi: 10.35889/jutisi.v13i2.2168.

[13] I. P. Rahayu, A. Fauzi, and J. Indra, "Analisis Sentimen Terhadap Program Kampus Merdeka Menggunakan Naive Bayes Dan Support Vector Machine," *Jurnal Sistem Komputer dan Informatika (JSON)*, vol. 4, no. 2, p. 296, Dec. 2022, doi: 10.30865/json.v4i2.5381.

[14] F. N. Azzahra, T. Rohana, R. Rahmat, and A. R. Juwita, "Penerapan Metode Naive Bayes Dalam Klasifikasi Spam SMS Menggunakan Fitur Teks Untuk Mengatasi Ancaman Pada Pengguna," *Journal of Information System Research (JOSH)*, vol. 5, no. 3, pp. 873–880, 2024, Accessed: May 04, 2025. [Online]. Available: https://ejurnal.seminar-id.com/index.php/josh/article/view/5070

[15] N. Romadoni, A. Mutoi Siregar, D. S. Kusumaningrum, and T. Rohana, "Classification Model of Public Sentiments About Electric Cars Using Machine Learning," *Scientific Journal of Informatics*, vol. 11, no. 2, 2024, doi: 10.15294/sji.v11i2.1309.

[16] A. D. Sidik and A. Ansawarman, "Prediksi Jumlah Kendaraan Bermotor Menggunakan Machine Learning," *Formosa Journal of Multidisciplinary Research*, vol. 1, no. 3, pp. 559–568, Jul. 2022, doi: 10.55927/fjmr.v1i3.745.

[17] Y. A. Singgalen, "Analisis Sentimen Wisatawan terhadap Taman Nasional Bunaken dan Top 10 Hotel Rekomendasi Tripadvisor Menggunakan Algoritma SVM dan DT berbasis CRISP-DM," *Journal of Computer System and Informatics (JoSYC)*, vol. 4, no. 2, pp. 367–379, Feb. 2023, doi: 10.47065/josyc.v4i2.3092.

[18] A. M. Pravina, I. Cholisoddin, and P. P. Adikara, "Analisis sentimen tentang opini maskapai penerbangan pada dokumen twitter menggunakan algoritme support vector machine (svm)," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, vol. 3, no. 3, pp. 2789–2797, 2019, [Online]. Available: https://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/4793

[19] Ihsan Zulfahmi, "Analisis Sentimen Aplikasi PLN Mobile Menggunakan Metode Decission Tree," *Jurnal Penelitian Rumpun Ilmu Teknik*, vol. 3, no. 1, pp. 11–21, Dec. 2023, doi: 10.55606/juprit.v3i1.3096.

[20] N. Arib Fadhlurrohman, A. Primajaya, and A. Nugraha Dimyati, "Analisis Sentimen Terhadap Skema Student Loan Untuk Biaya Perguruan Tinggi Pada Twitter Menggunakan Algoritma Naïve Bayes," *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. 9, no. 2, pp. 2115–2123, Mar. 2025, doi: 10.36040/jati.v9i2.13001.