eISSN : 2621-4970
pISSN : 2621-4962

Volume 9 | No 2
Juli | **2026**

Analisis Sentimen Netizen Terhadap Kenaikan Pajak Pertambahan Nilai (PPN) Di Media Sosial X: Perbandingan Algoritma Naive Bayes Dan C4.5

Achmad Munawar[1], Erick Harlest Budi Raharjo[2], Hafiz Fadhilah[3], Rafid Novriansyah[4]
Fakultas Teknologi Informasi, Program Studi Sistem Informasi
Universitas Bina Sarana Informatika
Achmad.amw@bsi.ac.id

Abstrak

Kenaikan tarif Pajak Pertambahan Nilai (PPN) memicu dinamika opini publik di media sosial X yang penting dianalisis sebagai bahan evaluasi kebijakan fiskal. Meskipun demikian, sebagian besar penelitian terdahulu masih berfokus pada data kualitatif atau data kategorial tabular, sehingga terdapat celah penelitian dalam mengukur respons publik secara kuantitatif pada teks berdimensi tinggi (sparse text). Penelitian ini bertujuan menganalisis sentimen netizen serta membandingkan kinerja algoritma Naïve Bayes dan C4.5 dalam mengklasifikasikan opini tersebut. Dataset terdiri dari 369 tweet valid hasil crawling yang diolah melalui preprocessing dan pembobotan TF-IDF. Hasil penelitian menunjukkan sentimen kontra lebih dominan (305 tweet) dibandingkan pro (64 tweet). Berdasarkan pengujian dengan rasio train-test split 70:30, Naïve Bayes mencatatkan akurasi sebesar 92,73%, mengungguli C4.5 yang memperoleh 86,36%. Hal ini membuktikan Naïve Bayes lebih efektif untuk klasifikasi teks berdimensi tinggi pada isu kebijakan publik. Secara praktis, dominasi sentimen kontra memberikan masukan bagi pemangku kebijakan dalam merumuskan strategi komunikasi publik. Secara akademis, temuan ini memperkuat bukti keunggulan algoritma probabilistik pada klasifikasi teks. Kendati demikian, penelitian ini masih terbatas pada ukuran dataset yang kecil (369 tweet), klasifikasi biner, dan evaluasi tanpa k-fold cross-validation, sehingga disarankan bagi penelitian selanjutnya untuk memperluas volume data dan menerapkan validasi silang.

Kata kunci: Analisis sentimen, Pajak Pertambahan Nilai (PPN), media sosial X, Naïve Bayes, C4.5, Text mining, Klasifikasi sentimen.

Abstract

The increase in Value Added Tax (VAT) rates has sparked public opinion dynamics on social media X, making its analysis essential for evaluating fiscal policy. Nevertheless, most previous studies still focused on qualitative data or tabular categorical data, leaving a research gap in quantitatively measuring public responses on high-dimensional text (sparse text). This study aims to analyze netizens' sentiments and compare the performance of Naïve Bayes and C4.5 algorithms in classifying these opinions. The dataset comprises 369 valid tweets collected through crawling, processed via preprocessing techniques and TF-IDF weighting. The findings reveal that opposing sentiment was dominant (305 tweets) compared to supportive sentiment (64 tweets). Based on testing with a 70:30 train-test split ratio, Naïve Bayes achieved an accuracy of 92.73%, outperforming C4.5 which reached 86.36%. This demonstrates that Naïve Bayes is more effective for high-dimensional text classification in public policy contexts. Practically, the dominance of opposing sentiment provides valuable input for policymakers in formulating public communication strategies. Academically, these findings strengthen empirical evidence supporting the superiority of probabilistic algorithms in text classification. Nevertheless, this study is limited by a small dataset size (369 tweets), binary classification, and evaluation without k-fold cross-validation, suggesting that future research expand the data volume and apply cross-validation.

Keywords: ***Sentiment analysis, Value Added Tax (VAT), social media X, Naive Bayes, C4.5, text mining, sentiment classification.***

PENDAHULUAN

Perkembangan media sosial telah membawa perubahan signifikan dalam pola komunikasi masyarakat, terutama dalam menyampaikan opini terhadap kebijakan publik. X, yang sebelumnya dikenal sebagai Twitter, merupakan salah satu platform media sosial yang banyak digunakan oleh masyarakat Indonesia untuk mengekspresikan pandangan, tanggapan, dan kritik terhadap berbagai isu, termasuk kebijakan kenaikan tarif Pajak Pertambahan Nilai (PPN). Kebijakan tersebut menimbulkan berbagai respons di tengah masyarakat, baik yang bersifat pro maupun kontra, yang tercermin dalam berbagai unggahan, komentar, dan diskusi di ruang digital.

Tingginya volume data teks yang dihasilkan melalui media sosial menjadikan analisis sentimen sebagai pendekatan yang relevan untuk mengidentifikasi kecenderungan opini publik secara otomatis dan sistematis. Analisis sentimen merupakan salah satu cabang text mining yang memanfaatkan teknik machine learning untuk mengklasifikasikan opini ke dalam kategori tertentu, seperti positif, negatif, atau netral. Dalam penelitian ini, klasifikasi sentimen dilakukan dengan menggunakan algoritma Naive Bayes dan C4.5 (Decision Tree), dengan metode TF-IDF sebagai teknik pembobotan fitur.

Naive Bayes dikenal sebagai algoritma probabilistik yang sederhana dan memiliki performa yang baik dalam klasifikasi data teks. Di sisi lain, C4.5 merupakan algoritma berbasis pohon keputusan yang mampu menghasilkan aturan klasifikasi secara hierarkis dan mudah dipahami. Perbandingan kedua algoritma tersebut bertujuan untuk mengetahui model yang paling optimal dalam mengklasifikasikan sentimen netizen terhadap kebijakan kenaikan PPN pada media sosial X.

Beberapa penelitian terdahulu telah menerapkan machine learning dan text mining dalam mengkaji opini publik serta evaluasi kebijakan. Kajian analisis sentimen pada platform media sosial X (sebelumnya Twitter) mengenai isu kebijakan publik telah dibahas oleh Faisol & Norsain (2023) melalui pendekatan netnografi terkait kenaikan tarif PPN 11%, serta Safira dkk. (2023) yang menganalisis persepsi masyarakat terhadap layanan finansial digital. Lebih lanjut, pengkajian dampak ekonomi dan stabilitas makro dari kebijakan fiskal pemerintah seperti PPN dilaporkan secara kualitatif dan kuantitatif oleh Kharisma & Furqon (2023) serta Subur dkk. (2025).

Dalam ranah ekstraksi fitur dan klasifikasi teks, metode pembobotan Term Frequency-Inverse Document Frequency (TF-IDF) terbukti efektif dalam menyaring kata-kata penting pada dokumen berdimensi tinggi, sebagaimana ditunjukkan oleh Das et al. (2023) dan Utami dkk. (2022). Terkait efektivitas algoritma pembelajar (classifier), perbandingan antara probabilistic classifier berbasis Naïve Bayes dan decision tree berbasis C4.5 secara konsisten dikaji dalam berbagai bidang, seperti prediksi penyakit (Kohsasih & Situmorang, 2022; Kurniati, 2023), data mining umum (Kurniawan dkk., 2018), klasifikasi status gizi (Lasarudin dkk., 2022), hingga penerimaan bantuan sosial (Fitriani, 2020).

Di samping itu, algoritma jaringan saraf tiruan (ANN) juga diteliti oleh Yusliani dkk. (2023) sebagai metode alternatif klasifikasi sentimen.

Kebaruan (*novelty*) dari penelitian ini terletak pada integrasi pemodelan klasifikasi komparatif antara algoritma probabilistik Naïve Bayes dan algoritma pohon keputusan C4.5 berbasis pembobotan TF-IDF yang secara khusus diterapkan pada isu kebijakan fiskal terkini di Indonesia, yakni kenaikan tarif PPN pada platform media sosial X. Sementara sebagian besar penelitian terdahulu berfokus pada data kualitatif atau klasifikasi teks pada domain umum, penelitian ini mengisi celah (*research gap*) dengan menyajikan evaluasi empiris kuantitatif yang mengukur tingkat akurasi, presisi, dan recall terhadap dinamika opini publik mengenai kebijakan PPN. Hasil komparasi ini tidak hanya memberikan kontribusi teoritis mengenai performa algoritma pada data teks berdimensi tinggi (sparse text), tetapi juga memberikan implikasi praktis berupa pemetaaan rasio sentimen pro dan kontra sebagai masukan berbasis data (data-driven) bagi pemangku kebijakan fiskal pemerintah.

Penelitian ini diharapkan dapat memberikan gambaran empiris mengenai persepsi masyarakat terhadap kebijakan pemerintah serta menjadi referensi dalam pengembangan model analisis sentimen berbasis teks, khususnya pada konteks kebijakan publik.

Rumusan Masalah

Dalam menganalisis sentimen netizen terhadap kebijakan kenaikan Pajak Pertambahan Nilai (PPN) di media sosial X yang mampu:

1. Bagaimana sentimen netizen terhadap kebijakan kenaikan Pajak Pertambahan Nilai (PPN) di media sosial X.
2. Bagaimana kinerja algoritma Naive Bayes dalam mengklasifikasikan sentimen netizen terhadap kenaikan PPN.
3. Bagaimana kinerja algoritma C4.5 dalam mengklasifikasikan sentimen netizen terhadap kenaikan PPN.
4. Algoritma manakah yang memberikan performa terbaik berdasarkan nilai accuracy, precision, dan recall.

Tujuan Penelitian

Tujuan dari penelitian ini adalah sebagai berikut:

1. Menganalisis sentimen netizen terhadap kebijakan kenaikan Pajak Pertambahan Nilai (PPN) pada media sosial X.
2. Mengolah data tweet terkait kenaikan PPN melalui proses crawling, preprocessing, dan pembobotan TF-IDF.
3. Menerapkan algoritma Naive Bayes dan C4.5 untuk klasifikasi sentimen.
4. Membandingkan performa kedua algoritma berdasarkan accuracy, precision, dan recall.
5. Menentukan algoritma yang paling efektif dalam analisis sentimen terkait kebijakan kenaikan PPN di media sosial X.

Manfaat Penelitian

Manfaat dari Tujuan yang akan dicapai dari penelitian adalah sebagai berikut:

1. Manfaat Teoritis
   Penelitian ini diharapkan dapat memberikan kontribusi terhadap pengembangan ilmu data mining dan

text mining, khususnya penerapan algoritma Naive Bayes dan C4.5 dalam analisis sentimen pada media sosial.

2. Manfaat Praktis
   Hasil penelitian dapat menjadi sumber informasi bagi pemerintah dan pemangku kebijakan dalam memahami persepsi masyarakat terhadap kebijakan kenaikan PPN berdasarkan opini yang berkembang di media sosial X.
3. Manfaat Akademis
   Penelitian ini dapat menjadi referensi bagi penelitian selanjutnya yang membahas analisis sentimen dan perbandingan algoritma klasifikasi teks.

Tinjauan Pustaka

Merujuk penelitian terdahulu dari penulis/peneliti sebelumnya terdahulu menunjukkan beragam pendekatan dalam menganalisis opini publik serta evaluasi kebijakan berbasis data digital. Dalam konteks evaluasi isu kebijakan fiskal di Indonesia, studi Faisol & Norsain (2023) mendokumentasikan dinamika opini publik terhadap kenaikan tarif PPN 11% secara mendalam. Namun, penelitian tersebut mengandalkan pendekatan kualitatif netnografi yang memiliki keterbatasan subjektivitas dan kurang efisien untuk mengolah volume data opini berukuran besar secara otomatis (automated classification). Laporan makro statistik oleh Kemp (2024) memang menegaskan besarnya penetrasi pengguna media sosial X di Indonesia, tetapi laporan tersebut bersifat lanskap demografis umum tanpa melakukan pemodelan komputasional atau pemetaan polaritas sentimen spesifik. Sejalan dengan hal itu, studi komunikasi publik oleh Adistian (2021) mengonfirmasi efektivitas platform X sebagai media penyampaian aspirasi masyarakat, meskipun belum mengintegrasikan teknik text mining atau machine learning untuk ekstraksi informasi terstruktur.

Pada ranah klasifikasi teks opini media sosial, penelitian Yusliani dkk. (2023) serta Al-Qablan et al. (2023) membuktikan efektivitas metode pembelajaran mesin dan deep learning dalam mengenali pola pola bahasa alami yang kompleks. Meskipun demikian, model berbasis Artificial Neural Network (ANN) memerlukan kompleksitas komputasi yang tinggi, resource yang besar, serta hyperparameter tuning yang rumit jika dibandingkan dengan algoritma klasifikasi konvensional. Di sisi lain, beberapa studi komparatif telah mengevaluasi kinerja algoritma Naïve Bayes dan C4.5 dalam berbagai domain, seperti penentuan bantuan sosial (Fitriani, 2020), prediksi penyakit (Kohsasih & Situmorang, 2022), dan kriteria medis (Kurniati, 2023). Akan tetapi, mayoritas pengujian tersebut dilakukan pada data terstruktur berjenis numerik atau kategorial (tabular). Terdapat celah analisis kritis mengenai bagaimana algoritma berbasis probabilistik (Naïve Bayes) dan pohon keputusan (C4.5) saling berkompetisi ketika diterapkan pada matriks teks tidak terstruktur yang berdimensi tinggi (high-dimensional sparse text), khususnya pada isu kebijakan fiskal PPN di media sosial X dengan pengayaan fitur pembobotan TF-IDF.

LANDASAN TEORI

Media sosial seperti Twitter (Adistian, 2021; Kemp, 2024) telah menjadi ruang interaktif bagi masyarakat dalam menyampaikan opini, termasuk terhadap kebijakan publik. Besarnya volume data

opini yang tidak terstruktur mendorong perlunya pendekatan text mining untuk memahami persepsi publik secara sistematis (Utami dkk., 2022). Analisis sentimen menjadi salah satu metode penting dalam mengklasifikasikan opini publik, baik positif maupun negatif (Novi Yusliani dkk., 2023).

Kebijakan kenaikan tarif Pajak Pertambahan Nilai (PPN) dari 10% menjadi 11% menimbulkan berbagai respons masyarakat. Penelitian Faisol & Norsain (2023), Kharisma & Furqon (2023), serta Subur dkk. (2025) menunjukkan adanya dampak sosial, ekonomi, dan inflasi yang mempengaruhi persepsi publik. Wibowo (2021) menekankan implementasi kenaikan tarif PPN pasca UU No. 7 Tahun 2021 pada dunia usaha, sehingga analisis sentimen publik menjadi relevan untuk mengevaluasi penerimaan kebijakan fiskal.

Menurut Al-Qablan et al. (2023), Analisis sentimen merupakan proses pengumpulan, identifikasi, dan klasifikasi opini, pemikiran, maupun kesan seseorang terhadap suatu topik ke dalam polaritas sentimen tertentu seperti positif, negatif, atau netral.

TF-IDF merupakan metode pembobotan fitur yang mengukur pentingnya suatu kata berdasarkan frekuensi kemunculannya pada dokumen tertentu dan kelangkaannya dalam keseluruhan corpus dokumen, Das, M., Selvakumar, K., & Alphonse, P.J.A. (2023).

Naive Bayes merupakan algoritma klasifikasi probabilistik yang menggunakan Teorema Bayes dengan asumsi independensi antar fitur dan banyak digunakan dalam klasifikasi data teks maupun analisis sentimen, Wardhani, I. P., Chandra, Y. I., & Yusra, F. (2023)

Algoritma C4.5 merupakan pengembangan dari ID3 yang membangun pohon keputusan dengan memanfaatkan nilai *Gain Ratio* dalam pemilihan atribut terbaik untuk proses klasifikasi. Kharisma, N., & Furqon, I. K. (2023) Proses pembentukan pohon keputusan pada C4.5 dilakukan secara rekursif *top-down (divide-and-conquer)* melalui evaluasi kecenderungan atribut. Mekanisme pembentukan pohon keputusan, Entropy, dan Gain Ratio.Entropy (Ketidakmurnian Data): C4.5 mengukur homogenitas atau tingkat ketidakmurnian (impurity) dari suatu kumpulan data (S) menggunakan konsep Entropy. Gain Ratio: Algoritma ID3 cenderung mengalami bias terhadap atribut yang memiliki banyak variasi nilai. Untuk mengatasi kelemahan tersebut, C4.5 memanfaatkan Gain Ratio sebagai kriteria pemilihan atribut pemisah (split attribute) terbaik.

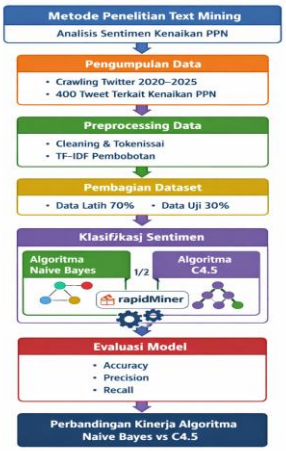
Dalam ranah klasifikasi sentimen, algoritma Naive Bayes dan C4.5 banyak digunakan karena efektivitasnya dalam mengolah data teks. Fitriani (2020), Kohsasih & Situmorang (2022), Kurniati Pratami (2023), Kurniawan dkk. (2018), serta Lasarudin dkk. (2022) membandingkan kedua algoritma dalam berbagai konteks, mulai dari kelayakan bantuan sosial hingga deteksi penyakit dan status gizi. Safira dkk. (2023) menegaskan efektivitas Naive Bayes dalam klasifikasi opini publik terhadap layanan finansial digital. Secara umum, hasil penelitian menunjukkan bahwa Naive Bayes cenderung lebih unggul dalam akurasi klasifikasi teks dibandingkan C4.5, meskipun keduanya memiliki kelebihan masing-masing.

Faisol dan Norsain (2023) menggunakan pendekatan etnografi untuk memahami perspektif netizen terhadap kenaikan tarif PPN 11%. Penelitian ini menyoroti opini publik di media sosial, memperlihatkan bagaimana kebijakan fiskal memicu respons beragam yang dapat dianalisis secara sistematis melalui data digital.

Kemp (2024) dalam laporan Digital 2024: Indonesia menyajikan data komprehensif mengenai tren penggunaan internet dan media sosial di Indonesia. Informasi ini relevan sebagai landasan penelitian sentimen publik, karena menggambarkan perilaku digital masyarakat yang menjadi sumber utama data opini.

METODE PENELITIAN

Metode yang digunakan dalam terdiri dari enam tahapan utama:



Gambar 1. Model Penelitian

a. Pendekatan Penelitian
   Penelitian ini menggunakan pendekatan kuantitatif berbasis text mining untuk menganalisis sentimen netizen terhadap kebijakan kenaikan Pajak Pertambahan Nilai (PPN) dari 10% menjadi 11%. Pendekatan kuantitatif dipilih karena penelitian berfokus pada pengolahan data numerik, pengukuran akurasi algoritma, serta perbandingan performa metode klasifikasi.
b. Data Penelitian
   Data penelitian berupa 400 tweet yang dikumpulkan melalui proses crawling menggunakan kata kunci terkait kenaikan PPN dalam rentang waktu 2020–2025. Setelah melalui tahapan preprocessing, diperoleh 369 *tweet valid* yang siap dianalisis, terdiri dari sentimen pro dan kontra sesuai hasil pembersihan data.
c. Tahapan *Preprocessing*
   Proses *preprocessing* dilakukan untuk memastikan kualitas data sebelum analisis. Tahapan meliputi: Penghapusan *URL*, mention, hashtag, dan simbol. Tokenisasi kata, Normalisasi huruf, Penghapusan stopwords. Representasi fitur dilakukan menggunakan metode TF-IDF (Term Frequency–Inverse Document Frequency) agar setiap kata memiliki bobot sesuai tingkat kepentingannya dalam dokumen.
d. Pembagian Dataset
   Dataset dibagi menjadi dua bagian: Data latih (70%) untuk membangun model klasifikasi. Data uji (30%) untuk mengukur performa model. Pembagian ini mengikuti praktik umum dalam penelitian berbasis machine learning untuk menghindari overfitting dan memastikan validitas hasil.
e. Algoritma Klasifikasi
   Proses klasifikasi dilakukan menggunakan dua algoritma: Naive Bayes, algoritma probabilistik yang sederhana namun efektif dalam mengklasifikasikan teks. C4.5, algoritma berbasis pohon keputusan yang mampu menangani data dengan

atribut kompleks. Kedua algoritma dijalankan menggunakan perangkat lunak RapidMiner, yang mendukung proses data mining dan analisis sentimen secara terintegrasi.

f. Evaluasi Model

Evaluasi performa model dilakukan menggunakan metrik: Accuracy tingkat ketepatan klasifikasi secara keseluruhan. Precision ketepatan klasifikasi terhadap sentimen tertentu. Recall kemampuan model dalam mendeteksi seluruh data sentimen yang relevan. Metrik ini digunakan untuk membandingkan kinerja Naive Bayes dan C4.5, sehingga dapat ditentukan algoritma yang lebih efektif dalam analisis sentimen publik terhadap kebijakan kenaikan PPN.

Proses pelabelan data sentimen dilakukan secara manual oleh tim peneliti yang bertindak sebagai anotator (annotators). Untuk menjaga konsistensi dan objektivitas penilaian, pelabelan mengacu pada pedoman (annotation guidelines) yang telah ditetapkan, di mana tweet dikategorikan ke dalam kelas Pro (berisi dukungan, apresiasi, atau tanggapan positif) dan Kontra (berisi kritik, penolakan, keberatan, atau sentimen negatif) terhadap kebijakan kenaikan PPN. Guna memastikan reliabilitas hasil pelabelan, proses anotasi melibatkan dua orang anotator independen. Apabila terjadi perbedaan penilaian (disagreement) antar-anotator terhadap suatu tweet, penyelesaian dilakukan melalui diskusi konsensus berdasarkan indikator kata kunci konteks pada pedoman pelabelan, atau melibatkan anotator ketiga (third annotator) untuk menentukan keputusan label akhir secara acuan silang (cross-validation).

Metode Pengumpulan Data

Untuk memperoleh data yang diperlukan dalam penelitian ini, penulis menggunakan teknik studi pustaka, observasi media sosial X, dan pengumpulan data tweet melalui proses crawling terkait isu kenaikan Pajak Pertambahan Nilai (PPN):

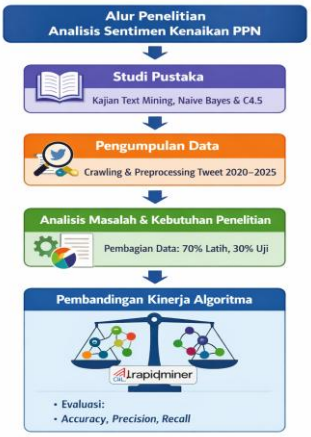
a. Teknik Studi Pustaka

Studi pustaka dilakukan dengan menelaah berbagai sumber ilmiah terkait analisis sentimen, algoritma Naive Bayes dan C4.5, serta kebijakan fiskal PPN. Referensi diambil dari jurnal nasional dan internasional yang relevan, seperti penelitian Fitriani (2020), Kohsasih & Situmorang (2022), dan Faisol & Norsain (2023). Kajian pustaka ini menjadi dasar teoritis dalam merancang model klasifikasi sentimen dan menentukan parameter evaluasi yang digunakan dalam penelitian berbasis text mining.

b. Teknik Pengamatan (Observation)

Metode observasi dilakukan dengan mengamati aktivitas dan interaksi netizen di media sosial X yang membahas kenaikan Pajak Pertambahan Nilai (PPN) dari 10% menjadi 11%. Observasi ini bertujuan untuk memahami pola komunikasi publik serta konteks sosial yang melatarbelakangi opini masyarakat. Data observasi mendukung proses *crawling tweet*, sehingga peneliti dapat menentukan kata kunci yang relevan dan memastikan bahwa data yang dikumpulkan mencerminkan sentimen nyata pengguna.

Kerangka Berpikir/Alur Penelitian



Gambar 2. Kerangka Pemikiran Pemecahan Masalah.

HASIL DAN PEMBAHASAN

Proses Pengolahan Data

Penelitian ini diawali dengan serangkaian proses pengolahan data mentah sebagai fondasi utama dalam analisis sentimen terhadap kebijakan kenaikan tarif Pajak Pertambahan Nilai (PPN) di media sosial X. Keseluruhan tahapan dirancang untuk menghasilkan data yang bersih, relevan, dan representatif sebelum diterapkan algoritma klasifikasi Naive Bayes dan C4.5. Proses pengolahan dimulai dari pengumpulan data, pembersihan, pelabelan sentimen, hingga transformasi teks menjadi fitur numerik yang siap diolah.

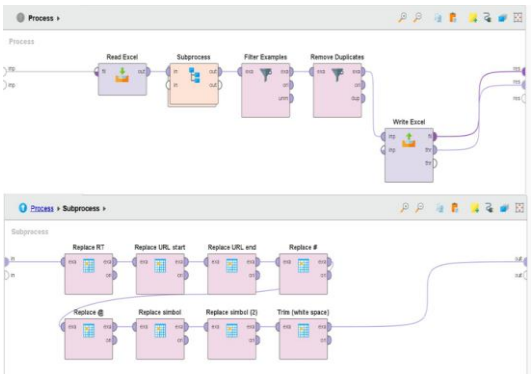
1. **Pengumpulan Data melalui Teknik Crawling**

   Data dalam penelitian ini dikumpulkan melalui teknik web crawling pada platform media sosial X dengan memanfaatkan layanan Google Collab, Node.js, dan alat bantu Tweet Harvest. Proses penelusuran difokuskan pada cuitan publik yang membahas isu kenaikan PPN dengan menggunakan kata kunci seperti “kenaikan PPN” dan “PPN 11%”. Periode pengambilan data mencakup rentang tahun 2020 hingga 2025, dan berhasil mengumpulkan sebanyak 400 tweet yang kemudian dijadikan sebagai dataset utama.

2. **Pembersihan Data (Data Cleaning)**

   Tahap pembersihan data menjadi langkah krusial untuk memastikan kualitas teks sebelum memasuki proses klasifikasi. Seluruh prosedur dilaksanakan secara sistematis menggunakan perangkat lunak RapidMiner. Proses diawali dengan mengimpor data mentah melalui operator Read Excel, dilanjutkan dengan serangkaian pemrosesan dalam subprocess yang mencakup penghapusan elemen-elemen non-informatif seperti retweet (RT), URL, hashtag (#), mention (@), simbol khusus, tanda baca, serta karakter non-alfanumerik yang tidak berkontribusi terhadap analisis sentimen.

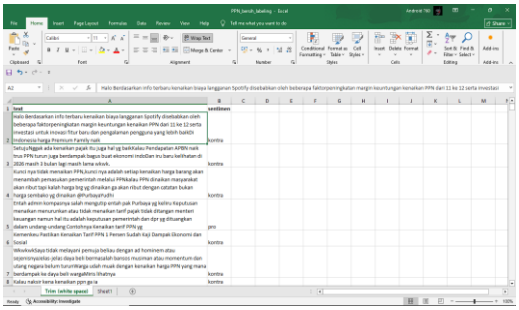
   Selain itu, dilakukan normalisasi teks dan penghapusan spasi berlebih (trim) untuk menjaga konsistensi token. Setelah pembersihan, langkah Filter Examples diterapkan untuk menyaring data kosong atau tidak relevan, sedangkan Remove Duplicates bertujuan menghilangkan duplikasi tweet agar tidak terjadi bias distribusi. Hasil akhir dari tahap ini disimpan melalui Write Excel sebagai dataset bersih yang siap untuk pelabelan dan pemodelan.



Gambar 3. Proses Cleaning Data.

3. **Pelabelan Sentimen**

Pada tahap pelabelan, sebanyak 400 tweet yang telah melalui proses seleksi diberikan label sentimen secara manual ke dalam dua kategori, yakni sentimen positif (pro) dan sentimen negatif (kontra). Proses ini bertujuan menyiapkan data terstruktur yang akan digunakan sebagai dasar pembelajaran mesin pada tahap klasifikasi.



Gambar 4. Proses Pelabelan Sentimen.

4. **Komposisi Dataset**

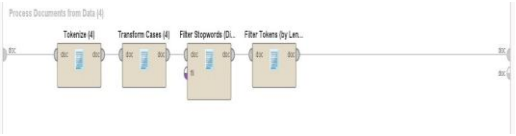
Setelah melalui seluruh rangkaian tahapan awal, dataset yang semula berjumlah 400 tweet menyusut menjadi 369 data yang lolos proses preprocessing. Hasil pelabelan menunjukkan bahwa sentimen kontra mendominasi dengan total 305 tweet, sementara sentimen pro hanya berjumlah 64 tweet. Ketimpangan distribusi ini menjadi perhatian pada tahap pemodelan guna mencegah terjadinya bias klasifikasi.

Table 1. Komposisi Dataset

Data Crawling	Sentimen Pro	Sentimen kontra	Data Akhir
400	64	305	369

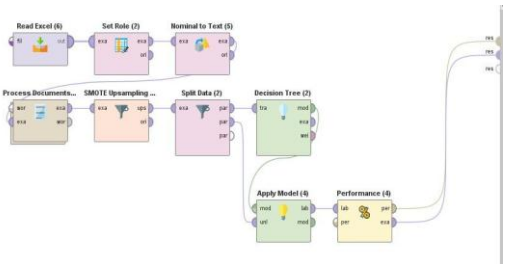
5. **Transformasi Dokumen Teks**

Untuk mengubah data teks menjadi representasi yang dapat diproses oleh algoritma klasifikasi, penelitian ini menerapkan serangkaian operator dalam RapidMiner, meliputi Tokenize, Transform Cases, Filter Stopwords, dan Filter Tokens (by Length).



Gambar 5. Tahap Transformasi Dokumen Teks

Pada tahap Tokenize, teks dipecah menjadi unit kata (token) yang berfungsi sebagai fitur analisis. Kemudian, Transform Cases digunakan untuk menyeragamkan huruf kapital, diikuti dengan Filter Stopwords untuk menghilangkan kata-kata umum yang tidak memiliki bobot opini. Terakhir, Filter Tokens (by Length) menyaring token berdasarkan panjang karakter sehingga hanya kata-kata yang bermakna yang dipertahankan.



Gambar 6. Tahap Transformasi klasifikasi Dokumen Teks

Rangkaian proses ini bertujuan meningkatkan kualitas representasi teks sehingga hasil klasifikasi sentimen lebih akurat dan terhindar dari noise

6. **Pembobotan Kata dengan TF-IDF**

Metode pembobotan kata yang digunakan adalah Term Frequency (TF) dan Term Frequency–Inverse Document Frequency (TF-IDF). TF mengukur seberapa sering suatu kata muncul dalam satu dokumen, sementara IDF mengurangi bobot kata yang sering muncul di seluruh dokumen dan sekaligus menekankan kata-kata yang bersifat spesifik pada dokumen tertentu. Hasil pembobotan TF-IDF inilah yang kemudian

dijadikan sebagai fitur masukan (input features) bagi algoritma Naive Bayes dan C4.5.

7. **Pengujian Model Klasifikasi**

Pengujian dilakukan untuk membandingkan kinerja kedua algoritma dalam mengklasifikasikan sentimen terhadap kebijakan PPN. Pada tahap ini, data hasil transformasi teks diimpor ke RapidMiner melalui operator Read Excel, lalu atribut label ditetapkan menggunakan Set Role. Proses konversi teks ke bentuk numerik dilanjutkan dengan Process Documents from Data yang memuat seluruh tahapan sebelumnya.

Pengujian model dilakukan dengan membagi dataset menggunakan rasio 70:30, di mana 70% (259 data) digunakan sebagai data latih dan 30% (110 data) sebagai data uji. Algoritma Naïve Bayes diterapkan untuk membangun model klasifikasi probabilistik, yang kemudian dievaluasi menggunakan confusion matrix untuk mengukur akurasi, presisi, dan recall.

Table 2. Akurasi Naive Bayes

	True Kontra	True Pro	Class Precision
Pred. Kontra	49	2	96,08%
Pred. Pro	6	53	89,83%
Class Recall	89,09%	96,36%	
Akurasi Keseluruhan: 92,73%			

$$Accuracy = \frac{TP+TK}{TP+TK+FP+FK} * 100\ \%$$

$$\frac{53+49}{110} * 100\% = 0.9273$$

$$= 92{,}73\%$$

Berdasarkan pengujian dengan rasio pembagian dataset 70% data latih (259 data) dan 30% data uji (110 data), model klasifikasi berhasil mencapai akurasi keseluruhan sebesar 92,73%. Dari total 110 data uji, model mampu memprediksi 102 data dengan benar dan hanya melakukan 8 kekeliruan. Secara rinci, model berhasil mengklasifikasikan 49 data True Kontra dan 53 data True Pro secara akurat. Sementara itu, kesalahan prediksi terjadi ketika 6 data kontra teridentifikasi sebagai pro, dan 2 data pro teridentifikasi sebagai kontra. Melalui hasil ini, model mencatatkan nilai Class Precision masing-masing sebesar 96,08% (Kontra) dan 89,83% (Pro), serta nilai Class Recall sebesar 89,09% (Kontra) dan 96,36% (Pro).

$$\text{Precision (pro)} = \frac{TP}{TP+FP} * 100\ \%$$

$$\frac{53}{53+6} * 100\% = 89.83\%$$

Sejalan dengan matriks evaluasi (confusion matrix) yang diperoleh, algoritma Naïve Bayes menunjukkan performa klasifikasi yang sangat baik dengan tingkat presisi rata-rata sebesar 92,96%. Model ini mencatatkan nilai presisi tertinggi pada kelas Kontra, yaitu sebesar 96,08%, sementara kelas Pro menghasilkan presisi sebesar 89,83%. Hasil tersebut mengindikasikan bahwa model memiliki tingkat validitas yang lebih tinggi dan lebih konsisten saat memprediksi sentimen atau kelas Kontra dibandingkan dengan kelas Pro.

$$\text{Recall (pro)} = \frac{TP}{TP+FK} * 100\ \%$$

$$= \frac{53}{53+2} * 100\% = 0{,}9636$$

$$= 96.36\%$$

Persentase Error Kontra = 100% - Recall Kontra.

100% - 89,09% = 10,91%

Hasil evaluasi berbasis confusion matrix menunjukkan bahwa algoritma Naïve Bayes memiliki kemampuan generalisasi yang solid, tercermin dari

capaian macro-average recall sebesar 92,73%. Tingkat sensitivitas tertinggi ditunjukkan oleh kelas Pro dengan nilai recall mencapai 96,36%, sebuah indikasi bahwa sistem mampu mengenali hampir seluruh representasi data Pro secara presisi. Di sisi lain, kelas Kontra mencatatkan recall sebesar 89,09%; gap minor ini mengisyaratkan bahwa sekitar 10,91% data aktual Kontra masih terdistribusi ke dalam kelas yang salah akibat misklasifikasi.

Pengujian model dilakukan melalui skema pembagian dataset dengan rasio 70:30, yang mengalokasikan 70% (259 data) sebagai data latih dan 30% (110 data) sebagai data uji. Model klasifikasi dibangun menggunakan algoritma C4.5, untuk kemudian dievaluasi kinerjanya menggunakan confusion matrix guna mengukur metrik akurasi, presisi, serta recall.

Table 3. Akurasi C4.5

	True Kontra	True Pro	Class Precision
Pred. Kontra	54	14	79,41%
Pred. Pro	1	41	97,62%
Class Recall	98,18%	74,55%	

$$\text{Accuracy} = \frac{TP+TK}{TP+TK+FP+FK} * 100\ \%$$

$$= \frac{41+54}{110} * 100\%$$

$$= 0.8636 = 86.\ 36\%$$

Pengujian model dilakukan menggunakan skema pembagian dataset berasio 70% untuk data latih (259 data) dan 30% untuk data uji (110 data). Dari total 110 data uji tersebut, evaluasi model menghasilkan tingkat akurasi keseluruhan sebesar 86,36%. Secara rinci, model berhasil mengklasifikasikan 54 data kontra dan 41 data pro dengan tepat, sehingga mengakumulasikan total 95 prediksi benar. Sementara itu, kekeliruan klasifikasi bersifat minor dengan detail 1 data kontra yang salah teridentifikasi sebagai pro, serta 14 data pro yang keliru dikenali sebagai kontra.

$$\text{Precision (pro)} = \frac{TP}{TP+FP} * 100\ \%$$

$$= \frac{41}{41+1} * 100\%$$

$$= 0.9762 = 97.62\%$$

Pengujian matriks evaluasi (confusion matrix) yang diperoleh, algoritma C4.5 menunjukkan performa klasifikasi dengan tingkat presisi rata-rata mencapai 88,52%. Kinerja terbaik secara signifikan terlihat pada kelas Pro, di mana nilai presisi menyentuh angka 97,62%. Capaian ini menandakan bahwa hampir seluruh data yang diprediksi sebagai Pro terbukti valid dan akurat. Sementara itu, kelas Kontra menghasilkan nilai presisi sebesar 79,41%, yang mengindikasikan bahwa masih terdapat sekitar 20,59% data yang keliru diklasifikasikan ke dalam kategori tersebut sebagai false positive.

$$\text{Recall (pro)} = \frac{TP}{TP+FK} * 100\ \%$$

$$= \frac{41}{41+14} * 100\%$$

$$= 0.7455 = 74.55\%$$

Berdasarkan matriks evaluasi (confusion matrix) yang diperoleh, algoritma C4.5 menunjukkan performa klasifikasi dengan tingkat presisi rata-rata mencapai 88,52%. Kinerja terbaik secara signifikan terlihat pada kelas Pro, di mana nilai presisi menyentuh angka 97,62%. Capaian ini menandakan bahwa hampir seluruh data yang diprediksi sebagai Pro terbukti valid dan akurat. Sementara itu, kelas Kontra menghasilkan nilai presisi sebesar 79,41%, yang mengindikasikan bahwa masih terdapat sekitar 20,59% data

yang keliru diklasifikasikan ke dalam kategori tersebut sebagai false positive. Persentase Error Kontra = 100% - Recall Kontra.
100% - 74,55% = 25,45%

Hasil evaluasi berbasis confusion matrix menunjukkan bahwa algoritma C4.5 memiliki kemampuan generalisasi yang cukup baik, tercermin dari capaian macro-average recall sebesar 86,37%. Tingkat sensitivitas tertinggi ditunjukkan oleh kelas Kontra dengan nilai recall yang sangat signifikan hingga mencapai 98,18%. Di sisi lain, kelas Pro mencatatkan recall yang lebih rendah yaitu sebesar 74,55%. Gap ini mengisyaratkan bahwa masih terdapat sekitar 25,45% data aktual Pro yang luput dari sistem dan keliru terdistribusi ke dalam kelas yang salah akibat misklasifikasi (false negative).

Meskipun model yang dibangun memberikan hasil yang optimal, penelitian ini memiliki beberapa keterbatasan, antara lain: (1) ukuran dataset yang dianalisis relatif kecil yaitu sebanyak 369 tweet valid; (2) klasifikasi sentimen terbatas pada dua kelas biner (pro dan kontra) tanpa mengakomodasi kelas netral; serta (3) evaluasi model hanya menggunakan satu skema pembagian dataset (70:30 train-test split) dan belum menerapkan pengujian validasi silang (k-fold cross-validation). Untuk penelitian selanjutnya, disarankan untuk memperluas volume data dari berbagai platform media sosial, menambahkan kategori sentimen yang lebih spesifik, serta menerapkan metode k-fold cross-validation guna meningkatkan reliabilitas, generalisasi, dan transparansi model.

KESIMPULAN

Berdasarkan hasil penelitian, sentimen netizen di media sosial X terhadap kebijakan kenaikan Pajak Pertambahan Nilai (PPN) secara signifikan didominasi oleh respons kontra (305 tweet) dibandingkan pro (64 tweet). Evaluasi performa model komparatif yang dilakukan setelah membagi dataset dengan rasio 70% data latih (259 tweet) dan 30% data uji (110 tweet) dari total 369 tweet valid menunjukkan bahwa algoritma Naïve Bayes mencatatkan performa terbaik dengan akurasi sebesar 92,73%, mengungguli algoritma C4.5 yang memperoleh akurasi sebesar 86,36%. Dengan demikian, algoritma Naïve Bayes terbukti lebih efektif dan optimal dalam mengklasifikasikan sentimen netizen terkait isu kenaikan PPN di media sosial X. secara praktis, tingginya respons kontra menjadi masukan penting bagi pemangku kebijakan dalam merumuskan strategi komunikasi publik dan mitigasi dampak kebijakan fiskal.

KETERBATASAN DAN SARAN

Meskipun memberikan hasil klasifikasi yang optimal, penelitian ini memiliki beberapa keterbatasan, yaitu: Ukuran dataset yang dikaji relatif kecil (369 tweet valid). Klasifikasi sentimen masih terbatas pada dua kelas biner (pro dan kontra) tanpa mengakomodasi kelas netral, serta Evaluasi performa model hanya mengandalkan satu skema pembagian data (*train-test split* 70:30) dan belum menerapkan pengujian validasi silang (*k-fold cross-validation*). Untuk penelitian selanjutnya, disarankan untuk memperluas volume dataset dari berbagai platform media sosial, mengakomodasi variasi kelas sentimen yang lebih spesifik, lanjut menerapkan teknik *k-fold cross-*

validation dan eksplorasi algoritma *ensemble/deep learning* guna meningkatkan generalisasi dan reliabilitas model.

UCAPAN TERIMA KASIH

Dengan penuh rasa syukur, penulis memanjatkan puji dan terima kasih ke hadirat Allah SWT atas limpahan rahmat dan kemudahan yang diberikan, sehingga penelitian ini dapat terselesaikan. Penulis menyampaikan apresiasi dan terima kasih yang mendalam kepada semua pihak yang telah meluangkan waktu, tenaga, dan pikiran untuk memberikan bimbingan serta arahan yang sangat berharga. Tidak lupa, penulis juga mengucapkan terima kasih kepada keluarga dan sahabat yang senantiasa menjadi sumber kekuatan melalui doa dan dukungannya. Ucapan khusus penulis tunjukan pula untuk diri sendiri, atas ketekunan dan kerja keras yang telah dijalani sepanjang proses penelitian ini.

Semoga segala bentuk bantuan dan kebaikan yang telah diberikan mendapat balasan terbaik dan menjadi amal jariyah yang tak terputus. Terima kasih.

DAFTAR PUSTAKA

Adistian, B. (2021). *Peran Twitter sebagai Media Informasi Pesan Dakwah Mahasiswa KPI UIN Raden Intan Lampung* (Skripsi, UIN Raden Intan Lampung).*http://repository.radenintan.ac.id/14589/*

Al-Qablan, T. A., Noor, M. H. M., & Khader, A. (2023). *A survey on sentiment analysis and its applications. Neural Computing and Applications*, 35, 22527–22547. *https://doi.org/10.1007/s00521-023-08941-x*

Das, M., Selvakumar, K., & Alphonse, P. J. A. (2023). *A comparative study on TF-IDF feature weighting method and its analysis using unstructured dataset*. arXiv preprint arXiv:2308.04037. *https://doi.org/10.48550/arXiv.2308.04037*

Faisol, M., & Norsain. (2023). Netnografi: *Perspektif Netizen terhadap Kenaikan Tarif PPN 11%.* Jurnal Analisis Aktual, 6(2), 167–182. *https://doi.org/10.22219/jaa.v6i2.24536*

Kemp, S. (2024). *Digital 2024*: Indonesia. DataReportal.*https://datareportal.com/reports/digital -2024-indonesia*

Kharisma, N., & Furqon, I. K. (2023). *Analisis Dampak Kenaikan Tarif Pajak Pertambahan Nilai (PPN)*. Jurnal Sahmiyya, 2, 295–303. *https://e-journal.uingusdur.ac.id/sahmiyya/article/view/1703*

Kohsasih, K. L., & Situmorang, Z. (2022). *Analisis Perbandingan Algoritma C4.5 dan Naïve Bayes dalam Memprediksi Penyakit Cerebrovascular.* Jurnal Teknologi Kesehatan, 9(1), 13–17. *https://doi.org/10.31294/inf.v9i1.11931*

Kurniati, P. I. S. W. (2023). *Perbandingan Algoritma C4.5 dan Naive Bayes dalam Mendeteksi Hipertensi di Puskesmas Banyubiru*. Jurnal Kesehatan Masyarakat, 2, 16–26. *https://ejournal.undip.ac.id/index.php /jkm*

Kurniawan, Y. I., Surakarta, U. M., & Bayes, N. (2018). *Comparison of Naive Bayes and C4.5 Algorithm in Data Mining*. Jurnal Teknologi Informasi dan Ilmu Komputer, 5(4), 455–464. *https://doi.org/10.25126/jtiik.201854965*

Lasarudin, A., Gani, H., & Tomayahu, M. (2022). *Perbandingan Metode Naïve Bayes dan C4.5 Klasifikasi Status Gizi Bayi Balita*. SPECTA Journal of Technology, 6(3), 273–283. *https://doi.org/10.35718/specta.v6i3.789*

Lestari, S., Yulmaini, Aswin, et al. (2022). *Implementation of the C4.5 Algorithm for Micro*, Small, and Medium Enterprises Classification. *https://doi.org/10.11591/ijece.v12i3.pp2859-2868*

Novi Yusliani, Armenia Yuhafi, Mastura Diana Marieska, & A. S. U. (2023). *Analisis Sentimen di Twitter Menggunakan Algoritma Artificial Neural Network*. JUPITER (Jurnal Penelitian Ilmu dan Teknologi Komputer), 15, 725–731. *https://journal.unwira.ac.id/index.php/JUPITER/article/view/2150*

Safira, A., Hasan, F. N., Timur, K. J., & Classifier, N. B. (2023). *Analisis Sentimen Masyarakat terhadap Paylater Menggunakan Metode Naive Bayes Classifier*. Jurnal Sistem Infomasi, 5(1), 59–70. *https://doi.org/10.31599/jiko.v5i1.2105*

Subur, Hikmayani, W. M. S. (2025). *Analisis Dampak Kenaikan Tarif Pajak Pertambahan Nilai (PPN) terhadap Masyarakat dan Inflasi di Indonesia*. Jurnal Ekonomi dan Kebijakan, 1(5), 205–210. *https://doi.org/10.61722/jrme.v1i5.3045*

Utami, N. W., Artana, M., & Akuntansi, S. I. (2022). *Text Mining dalam Analisis Sentimen Pembelajaran Daring di Masa Pandemi COVID-19 Menggunakan Algoritma K-Nearest Neighbor*. Jurnal Teknik Informatika, 4(2), 140–148. *https://doi.org/10.31599/ jti.v4i2.1120*

Wibowo, D. (2021). *Implementasi Kenaikan Tarif PPN Pasca UU No 7 Tahun 2021 pada Pengusaha Kena Pajak di Surabaya* (Skripsi, Universitas Wijaya Kusuma Surabaya).*http://repository.uwks.ac.id/id/eprint/5102*

Wardhani, I. P., Chandra, Y. I., & Yusra, F. (2023). *Application of the Naïve Bayes Classifier Algorithm to Analyze Sentiment for the Covid-19 Vaccine on Twitter in Jakarta*. *https://doi.org/10.31294/inf.v10i1.14205*