BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 6, No 1, December 2025 | Hal 469-478
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v6i1.932

Klasifikasi Sentimen Bitcoin Terhadap Komentar Di Aplikasi X Menggunakan Metode Decision Tree C4.5

Habibi Putra Indrizal, Fadhilah Syafria*, Elin Haerani, Yelvi Vitriani, Yusra

Fakultas Sains dan Teknologi, Program Studi Teknik Informatika, Universitas Islam Negeri Sultan Syarif Kasim Riau, Pekanbaru, Indonesia
E-mail: [1]12050110466@students.uin-suska.ac.id,[2,*]fadhilah.syafria@uin-suska.ac.id,
[3]elin.haerani@uin-suska.ac.id,[4]yelfi.vitriani@uin-suska.ac.id,[5]yusra@uin-suska.ac.id
Email Penulis Korespondensi: fadhilah.syafria@uin-suska.ac.id

Abstrak−Analisis sentimen merupakan salah satu metode penting untuk memahami persepsi pengguna terhadap aset kripto seperti Bitcoin, yang pergerakan harganya sangat dipengaruhi oleh opini publik. Penelitian ini bertujuan untuk mengklasifikasikan sentimen komentar pengguna pada platform X ke dalam dua kelas, yaitu positif dan negatif, menggunakan algoritma Decision Tree C4.5. Dataset yang digunakan berjumlah 5.000 komentar berbahasa Indonesia yang diperoleh melalui proses web scraping dan telah melalui tahapan preprocessing teks serta ekstraksi fitur menggunakan TF-IDF. Model dilatih menggunakan skema pembagian data 70% data latih dan 30% data uji. Hasil evaluasi menunjukkan bahwa model C4.5 memperoleh akurasi sebesar 78%. Pada kelas positif, model mencapai nilai recall yang sangat tinggi sebesar 0.99 dengan F1-score sebesar 0.83, yang menunjukkan kemampuan model yang sangat baik dalam mengenali komentar positif. Sebaliknya, pada kelas negatif diperoleh recall sebesar 0.51 dengan F1-score sebesar 0.67, meskipun precision-nya tinggi sebesar 0.97. Perbedaan performa antar kelas ini dipengaruhi oleh distribusi data yang tidak sepenuhnya seimbang, di mana jumlah komentar positif lebih dominan dibandingkan komentar negatif, sehingga model cenderung lebih sensitif terhadap kelas mayoritas. Secara keseluruhan, hasil penelitian menunjukkan bahwa algoritma Decision Tree C4.5 cukup efektif untuk klasifikasi sentimen Bitcoin berbahasa Indonesia, namun masih memiliki keterbatasan dalam mengenali kelas minoritas. Penelitian selanjutnya dapat mengkaji penerapan teknik penanganan data tidak seimbang atau algoritma yang lebih kompleks untuk meningkatkan keseimbangan performa antar kelas.

Kata kunci: Sentimen; Bitcoin; Decision Tree C4.5; TF-IDF; Klasifikasi Teks

Abstract-Sentiment analysis is an important method for understanding user perceptions of cryptocurrency assets such as Bitcoin, whose price movements are strongly influenced by public opinion. This study aims to classify user sentiment from comments posted on the X platform into two classes, namely positive and negative, using the Decision Tree C4.5 algorithm. The dataset consists of 5,000 Indonesian-language comments collected through a web scraping process and processed through text preprocessing and TF-IDF–based feature extraction. The model was trained using a 70% training data and 30% testing data split. The evaluation results show that the C4.5 model achieved an accuracy of 78%. For the positive class, the model obtained a very high recall of 0.99 with an F1-score of 0.83, indicating strong performance in identifying positive comments. In contrast, the negative class achieved a recall of 0.51 with an F1-score of 0.67, despite having a high precision of 0.97. The disparity in performance between classes is influenced by the data distribution, which is not fully balanced, with positive comments being more dominant than negative ones, causing the model to be more sensitive to the majority class. Overall, the results indicate that the Decision Tree C4.5 algorithm is sufficiently effective for Indonesian-language Bitcoin sentiment classification, although it still has limitations in recognizing the minority class. Future research may explore the application of data imbalance handling techniques or more advanced algorithms to improve the balance of classification performance across classes.

Keywords: Sentiment; Bitcoin; Decision Tree C4.5; TF-IDF; Text Classification

1. PENDAHULUAN

Perkembangan teknologi digital telah mengubah cara masyarakat dalam berinteraksi dan menyampaikan opini melalui media sosial. X menjadi salah satu platform yang banyak digunakan untuk membicarakan isu-isu ekonomi digital, termasuk Bitcoin[1]. Karena Bitcoin tidak diatur oleh satu pihak atau lembaga tertentu, mata uang digital ini menimbulkan berbagai pendapat di masyarakat, ada yang optimis melihat peluangnya, namun ada juga yang khawatir dengan naik-turunnya harga serta risiko yang ada. Setiap harinya, komentar terkait Bitcoin dipublikasikan dalam jumlah yang sangat besar, dengan gaya bahasa yang singkat, informal, dan bervariasi. Kondisi tersebut menjadikan pemantauan dan analisis sentimen secara manual tidak memungkinkan secara manusiawi (humanly impossible), sehingga pendekatan otomatis sangat diperlukan[2].

Untuk mengatasi permasalahan tersebut, analisis sentimen berbasis machine learning dapat digunakan sebagai solusi otomatis dalam mengolah data teks dalam jumlah besar. Pendekatan ini memungkinkan pengelompokan opini pengguna ke dalam kategori sentimen tertentu, seperti positif dan negatif, secara lebih objektif. Berbagai algoritma klasifikasi telah diterapkan dalam analisis sentimen[3], namun tidak semua algoritma mudah dipahami cara kerjanya, terutama algoritma yang bersifat black-box. Salah satu algoritma yang relevan adalah *Decision Tree*.

Algoritma Decision Tree, khususnya C4.5[4], dipilih dalam penelitian ini karena memiliki struktur model yang sederhana dan mudah diinterpretasikan. Algoritma ini mampu menangani data teks hasil transformasi numerik serta menghasilkan aturan klasifikasi yang jelas dalam bentuk pohon keputusan. Keunggulan tersebut menjadikan C4.5 lebih transparan dibandingkan metode lain seperti *Naive Bayes*[5], *Support Vector Machine* atau *Neural Networks*. Untuk meningkatkan kualitas representasi data teks, penelitian ini menerapkan tahapan text preprocessing dan feature extraction menggunakan Term Frequency–Inverse Document Frequency (TF-IDF).

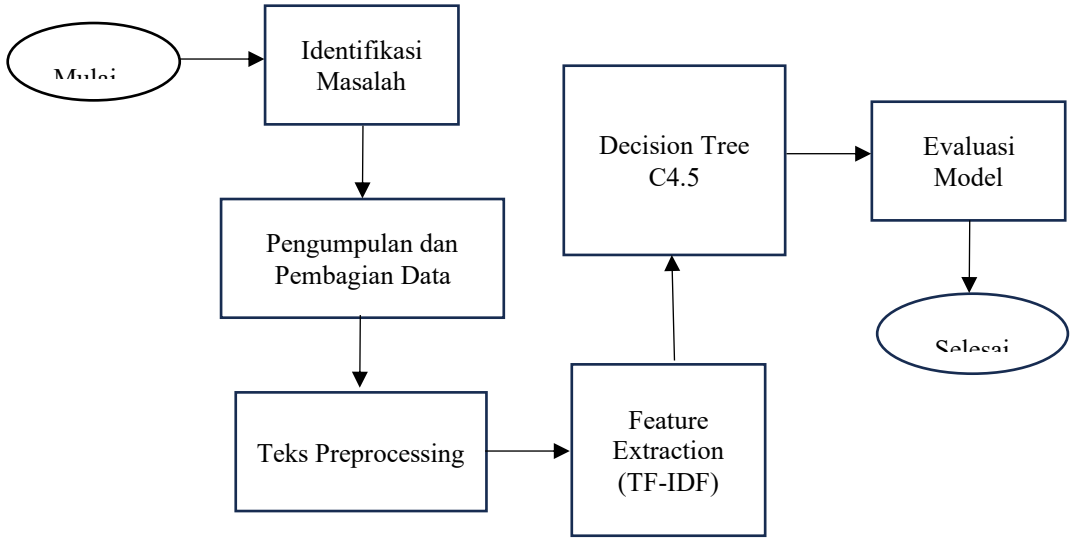
BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 6, No 1, December 2025 | Hal 469-478
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v6i1.932

Beberapa penelitian terdahulu menunjukkan bahwa algoritma C4.5 memiliki kinerja yang baik dalam tugas klasifikasi sentimen pada data media sosial[6]. Namun, sebagian besar penelitian tersebut masih berfokus pada topik umum atau menggunakan platform selain X, serta belum banyak yang secara khusus mengkaji analisis sentimen berbahasa Indonesia dengan topik Bitcoin[7]. Hal ini menunjukkan adanya celah penelitian (research gap) yang perlu dikaji lebih lanjut, terutama terkait pemanfaatan algoritma Decision Tree C4.5 untuk menganalisis sentimen publik terhadap Bitcoin di platform X.

Berdasarkan latar belakang tersebut, penelitian ini bertujuan untuk menganalisis sentimen publik terhadap Bitcoin melalui komentar pengguna di platform X dengan menggunakan algoritma Decision Tree[8]. Dataset yang digunakan terdiri dari 5.000 komentar berbahasa Indonesia yang diperoleh melalui proses web scraping. Fokus penelitian meliputi klasifikasi sentimen ke dalam kategori positif dan negatif, evaluasi kinerja model menggunakan metrik akurasi, presisi, recall, dan F1-score, serta identifikasi kata-kata yang berpengaruh dalam proses klasifikasi. Hasil penelitian ini diharapkan dapat memberikan kontribusi akademik dalam pengembangan kajian analisis sentimen berbahasa Indonesia serta memberikan manfaat praktis bagi investor, analis pasar, dan pihak terkait dalam memahami tren opini publik terhadap Bitcoin.

2. METODOLOGI PENELITIAN

Untuk mempermudah pemahaman alur penelitian yang dilakukan, maka disusun tahapan penelitian secara sistematis mulai dari tahap awal hingga tahap akhir. Tahapan ini mencakup proses identifikasi masalah, pengumpulan dan pembagian data, prapemrosesan teks, ekstraksi fitur menggunakan TF-IDF, proses klasifikasi menggunakan algoritma Decision Tree C4.5, hingga evaluasi model. Alur tahapan penelitian tersebut dapat dilihat pada Gambar 1.



Gambar 1. Tahapan penelitian

Berdasarkan Gambar 1, penelitian diawali dengan tahap identifikasi masalah untuk menentukan tujuan dan ruang lingkup penelitian. Selanjutnya dilakukan pengumpulan data yang kemudian dibagi menjadi data latih dan data uji. Data teks yang diperoleh melalui proses pengumpulan selanjutnya diproses pada tahap prapemrosesan teks untuk membersihkan dan menyiapkan data. Tahap berikutnya adalah ekstraksi fitur menggunakan metode TF-IDF untuk mengubah data teks menjadi representasi numerik. Fitur yang dihasilkan kemudian digunakan sebagai input pada algoritma Decision Tree C4.5 untuk melakukan klasifikasi sentimen. Tahap terakhir adalah evaluasi model untuk mengukur kinerja klasifikasi yang dihasilkan sebelum penelitian dinyatakan selesai.

2.1 Identifikasi Masalah

Penelitian ini bertujuan untuk mengklasifikasikan komentar pengguna mengenai Bitcoin menjadi sentimen positif atau negatif secara otomatis dan akurat dengan memanfaatkan algoritma machine learning, khususnya metode Decision Tree[9]. Peningkatan jumlah komentar di Aplikasi X yang memuat opini publik tentang Bitcoin, namun cenderung singkat, informal, dan bervariasi bahasanya, menjadi alasan perlunya pendekatan otomatis dalam proses analisis.

2.2 Evaluasi Dataset

a. Pengumpulan data
Data pada penelitian ini diperoleh melalui proses webscraping dari platform X. Proses webscraping dilakukan dengan memanfaatkan kata kunci “Bitcoin” untuk menangkap berbagai komentar publik. Seluruh data yang diperoleh merupakan data publik yang dapat diakses secara umum. Setelah proses pengumpulan, data kemudian dilakukan pembersihan awal(filtering) untuk memastikan hanya komentar berbahasa Indonesia yang digunakan dalam

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 6, No 1, December 2025 | Hal 469-478
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v6i1.932

penelitian. Dataset ini selanjutnya diproses pada tahap text preprocessing dan digunakan sebagai dasar analisis sentimen dengan metode Decision Tree dalam mengklasifikasikan sentimen positif maupun negatif terkait Bitcoin. Tabel 1 berikut menyajikan sebagian hasil data mentah yang diperoleh dari proses webscraping.

Tabel 1. Hasil Webscraping

No	full_text	id_str	created_at	username	user_id_str	...	source
1	Berita Bitcoin: 3 Alasan Kenapa BTC Masih Akan Tetap Turun di Bawah US$16.000 #bitcoin #kripto https://t.co/HU0DlFiV0a	1.61E+18	2023-01-01 08:09:13+00:00	cintabitcoin	84821316	...	cintabitcoin
2	@cryptondo Kapan bitcoin turun ke bawah lagi testing ke harga 10	1.61E+18	2023-01-02 02:49:08+00:00	miswar_riki	1.33E+18	...	miswar_riki
3	@DU09BTC #Bitcoin kalau turun ke support 14k bagus sekali... dan semoga rebound dari support ini	1.61E+18	2023-01-02 12:25:10+00:00	TraderCoinCrypt	1.45E+18	...	TraderCoinCrypt
...	...	...	...	...	...	...	...
5000	BTC nge-dump tanpa peringatan, mental horor semua #2242 sih!!	1.61E+18	2023-01-02 04:58:06+00:00	donicrypto	102967390	...	donicrypto

Tabel 1 hanya menampilkan beberapa contoh data mentah sebagai representasi dari total 5.000 komentar hasil webscraping. Data yang dikumpulkan merupakan komentar pengguna yang digunakan sebagai bahan analisis sentimen[10] dalam penelitian ini. Setiap baris pada tabel merepresentasikan satu entri komentar yang berhasil diambil melalui proses *webscraping*. Kolom *full_text* berisi isi lengkap teks komentar yang dipublikasikan oleh pengguna, termasuk tagar, *mention*, maupun tautan jika terdapat pada komentar asli. Kolom *id_str* dan *user_id_str* memperlihatkan id unik komentar dan id pengguna dalam format *string*, yang secara otomatis ditampilkan dalam bentuk notasi ilmiah karena panjangnya angka. Kolom created_at menunjukkan waktu ketika komentar tersebut diposting, berdasarkan penanda waktu yang diberikan langsung oleh sistem melalui API, sementara kolom *username* berisi nama pengguna yang mengunggah komentar tersebut. Kolom *source* menunjukkan platform atau aplikasi yang digunakan oleh pengguna saat membuat komentar, seperti *Twitter Web App*, *Android*, atau aplikasi pihak ketiga lainnya. Dataset ini menjadi dasar untuk tahapan *preprocessing* pada penelitian, seperti pembersihan teks, *normalisasi*, *tokenisasi*, *stopword removal*, serta *stemming*. Data mentah pada tabel tersebut menunjukkan variasi gaya penulisan yang cukup besar, seperti penggunaan bahasa informal, singkatan, serta slang, yang menjadikan proses preprocessing sangat penting untuk meningkatkan akurasi.

b. Labeling data
Setelah proses pengumpulan dan pembersihan data selesai, tahap berikutnya adalah pemberian label sentimen pada setiap data. Pelabelan dilakukan secara manual oleh tiga anotator dengan menelaah setiap komentar yang memuat kata kunci "Bitcoin". Setiap komentar kemudian diklasifikasikan ke dalam dua kategori sentimen, yaitu positif dan negatif, sementara data dengan sentimen netral dikeluarkan dari dataset untuk menjaga fokus dan kualitas proses klasifikasi. Apabila terjadi perbedaan penilaian antar anotator, dilakukan diskusi bersama hingga tercapai kesepakatan akhir. Proses pelabelan ini bertujuan untuk menghasilkan dataset berlabel yang berkualitas, sehingga dapat mendukung kinerja algoritma Decision Tree C4.5 dalam melakukan klasifikasi sentimen secara optimal.

c. Pembagian data
Setelah tahap preprocessing, dataset dibagi ke dalam tiga subset, yaitu data latih, data validasi, dan data uji. Data latih berfungsi untuk membangun model, sementara data validasi digunakan untuk menilai performa model selama proses pelatihan dan membantu menyesuaikan parameter agar tidak terjadi overfitting. Adapun data uji digunakan untuk mengevaluasi kemampuan model dalam mengklasifikasikan data baru yang tidak terlibat dalam proses pelatihan.

Pada penelitian ini digunakan rasio 70:30 dengan 70% data latih, 30% data uji. Selanjutnya 30% dibagi kembali dengan 15% data test dan 15% data validasi. Pembagian 70:30 dengan 30% pembagian 15%-15% dipilih karena secara teknis dianggap ideal untuk dataset berukuran sedang, porsi terbesar pada data latih memungkinkan model belajar lebih efektif, sementara proporsi validasi dan uji yang seimbang cukup untuk mengevaluasi performa dan generalisasi model secara akurat tanpa mengurangi jumlah data pelatihan secara signifikan. Dengan demikian, pembagian ini membantu memastikan model yang dihasilkan stabil dan reliabel.

2.3 Text Preprocessing

Text preprocessing bertujuan untuk membersihkan dan menyederhanakan teks agar dapat diolah lebih efektif oleh algoritma machine learning [11]. Proses ini mengubah teks mentah menjadi format yang lebih terstruktur. Adapun proses dari text preprocessing.

a. Case folding

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 6, No 1, December 2025 | Hal 469-478
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v6i1.932

Case folding adalah proses mengubah semua huruf dalam teks menjadi huruf kecil (lowercase).

b. Cleaning
   1. Penghapusan simbol, tanda baca dan emoji.
      Teks sering mengandung karakter yang tidak relevan dengan analisis seperti tanda baca, simbol, atau emoji. Tahap ini bertujuan menghilangkan tanda baca, simbol dan emoji.
   2. Penghapusan URL
      Penghapusan URL adalah tahap membersihkan teks dengan cara menghapus semua tautan (link/URL) yang biasanya muncul dalam data, terutama data dari media sosial seperti di platform x.
   3. Penghapusan angka
      Angka pada teks umumnya tidak diperlukan, kecuali dalam kasus tertentu (misalnya analisis harga). Pada preprocessing umum, angka dihapus agar tidak mengganggu analisis kata.
   4. Penghapusan karakter khusus
      Penghapusan karakter khusus adalah proses membersihkan teks dengan cara menghapus karakter yang bukan huruf alfabet.

c. Tokenisasi
   Tokenisasi adalah proses memecah kalimat menjadi kata-kata tunggal (token).

d. Stopword removal
   Stopword adalah kata-kata umum yang sering muncul tetapi tidak memiliki makna penting dalam analisis.

e. Stemming
   Stemming adalah proses mengubah kata menjadi bentuk dasar (root word).

2.4 Feature Extraction (TF-IDF)

Secara teknis, Term Frequency (TF) menunjukkan seberapa sering suatu kata muncul dalam dokumen, sementara Inverse Document Frequency (IDF) menunjukkan seberapa jarang kata itu muncul di dokumen lain[12]. Hasil dari TF kemudian dikalikan dengan IDF untuk menghasilkan skor TF-IDF bagi setiap kata di masing-masing dokumen[13]. (Rumus persamaan 1,2,3).

$$\text{TF: } tf_t = 1 + \log(tf_t) \tag{1}$$

$$\text{IDF: } idf_t = \log\left(\frac{D}{df_t}\right) \tag{2}$$

$$\text{TF – IDF: } W_{td} = tf_t \times idf_t \tag{3}$$

Dalam formulasi TF–IDF, simbol *t* merepresentasikan sebuah *term* atau kata, sedangkan *d* menunjukkan dokumen tempat kata tersebut muncul. Nilai tf_t menyatakan jumlah kemunculan kata *t* dalam dokumen *d*, yang kemudian ditransformasikan menggunakan fungsi logaritmik untuk menyeimbangkan pengaruh frekuensi kata. Sementara itu, *D* menunjukkan total jumlah dokumen dalam korpus, dan df_t menyatakan jumlah dokumen yang mengandung kata *t*, yang digunakan untuk menghitung nilai *idf* sebagai ukuran kelangkaan suatu kata dalam seluruh dokumen. Bobot akhir TF–IDF, yang dinotasikan sebagai *W*(t,d), diperoleh dari hasil perkalian antara *tf* dan *idf*, sehingga mencerminkan tingkat kepentingan suatu kata *t* dalam dokumen *d* relatif terhadap keseluruhan koleksi dokumen.

Setelah nilai TF-IDF diperoleh dari Persamaan (1), (2), dan (3), setiap term dalam kumpulan dokumen dikonversi menjadi bentuk numerik yang merefleksikan tingkat kepentingannya terhadap dokumen. Bobot-bobot tersebut kemudian disusun ke dalam sebuah vektor fitur berukuran besar, di mana setiap komponennya mewakili satu term unik pada dataset[14]. Vektor fitur ini digunakan sebagai input pada algoritma Decision Tree C4.5 dalam proses pembangunan model. Meskipun bobot TF-IDF bersifat kontinu, algoritma C4.5 mampu menangani atribut numerik dengan menentukan nilai ambang (threshold) optimal pada setiap fitur melalui perhitungan gain ratio, sehingga pemisahan data dilakukan berdasarkan titik potong bobot TF-IDF yang paling efektif untuk klasifikasi sentimen[15].

2.5 Decision Tree C4.5

2.5.1. Klasifikasi Decision Tree C4.5

Algoritma Decision Tree C4.5 merupakan teknik klasifikasi yang memanfaatkan struktur pohon keputusan untuk membangun model prediktif berdasarkan data yang telah memiliki label kelas[16]. Dalam proses konstruksi pohon, algoritma secara sistematis menentukan atribut yang paling mampu membedakan data antar kelas. Pemilihan atribut tersebut didasarkan pada dua ukuran utama, yaitu *Information Gain*, yang menggambarkan tingkat pengurangan ketidakpastian setelah pemisahan data, dan *Gain Ratio*, yaitu nilai *IG* yang telah dinormalisasi guna mengurangi kecenderungan pemilihan atribut dengan jumlah nilai unik yang besar[17]. Atribut dengan nilai *Gain Ratio* tertinggi selanjutnya digunakan sebagai titik pemisahan *split* pada setiap node dalam struktur pohon. Rumus yang digunakan pada bagian ini ditujukan oleh rumus persamaan 4,5,6,7,8.

a. *Entropy*
   Entropy adalah ukuran tingkat ketidakpastian atau ketidakteraturan dalam suatu himpunan data.

$$Entropy(S) = -\sum_{i=1}^{k} pi\ log_2(pi) \tag{4}$$

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 6, No 1, December 2025 | Hal 469-478
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v6i1.932

S merupakan himpunan data atau subset data yang sedang dihitung nilai entropinya, k menyatakan jumlah kelas yang terdapat dalam himpunan data S, sedangkan p_i adalah proporsi atau probabilitas data yang termasuk ke dalam kelas ke-i, yang digunakan sebagai dasar perhitungan tingkat ketidakpastian atau keragaman data.

$$p_i = \frac{jumlah\ data\ kelas\ ke{-}i}{total\ data} \tag{5}$$

$log_2(p_i)$= logaritma basis 2 dari probabilitas p_i, ∑= simbol penjumlahan untuk menjumlahkan seluruh nilai dari kelas 1 hingga kelas k

b. *Information gain*
Information Gain adalah ukuran penurunan ketidakpastian setelah data dibagi berdasarkan suatu atribut.

$$IG(S,A) = Entropy(S) - \sum_{v \epsilon A} \frac{|S_v|}{|S|} Entropy(S_v) \tag{6}$$

Information Gain $IG(S,A)$ merupakan nilai yang menunjukkan besarnya pengurangan entropi dari suatu atribut A terhadap himpunan data (S), di mana (Entropy(S)) adalah nilai entropi awal sebelum dilakukan proses pemisahan data. Atribut (A) digunakan sebagai dasar pembagian data ke dalam beberapa subset, dengan (v \in A) merepresentasikan setiap nilai atau kategori yang dimiliki oleh atribut tersebut, sehingga terbentuk subset (S_v) yang berisi data dengan nilai atribut (A = v). Banyaknya data dalam setiap subset dinyatakan dengan (|S_v|), sedangkan (|S|) menunjukkan jumlah total data dalam himpunan (S). Nilai (Entropy(S_v)) merupakan entropi dari masing-masing subset hasil pemisahan, dan simbol (\sum) digunakan untuk menjumlahkan seluruh entropi subset yang telah diberi bobot berdasarkan proporsi (|S_v|/|S|), sehingga diperoleh total entropi terpisah (weighted entropy).

c. *Split information*
Split Information adalah ukuran yang menggambarkan seberapa besar suatu atribut membagi data ke dalam beberapa subset.

$$SplitInfo(A) = -\sum_{v \epsilon A} \frac{|S_v|}{|S|} log_2\left(\frac{|S_v|}{|S|}\right) \tag{7}$$

SplitInfo(A) merupakan nilai split information dari atribut (A) yang digunakan untuk mengukur sejauh mana atribut tersebut membagi himpunan data (S) ke dalam beberapa subset. Atribut (A) adalah atribut yang sedang dievaluasi dalam proses pemisahan data, dengan (v∈A) merepresentasikan setiap nilai atau kategori yang dimiliki oleh atribut tersebut, sehingga terbentuk subset (Sv) sebagai bagian dari himpunan data (S) berdasarkan nilai atribut (A = v). Jumlah data pada masing-masing subset dinyatakan dengan (|Sv|), sedangkan (|S|) menunjukkan jumlah total data dalam himpunan (S). Nilai $log_2\left(\frac{|S_v|}{|S|}\right)$ merupakan logaritma basis dua dari proporsi data dalam subset (Sv) terhadap keseluruhan data (S), dan simbol (sum) digunakan untuk menjumlahkan seluruh nilai untuk setiap (v) pada atribut (A), sehingga menghasilkan ukuran pemisahan data yang lazim.

d. *Gain ratio*
Gain Ratio adalah nilai Information Gain yang sudah dinormalisasi oleh split information.

$$GainRatio(A) = \frac{IG(S,A)}{SplitInfo(A)} \tag{8}$$

$GainRatio(A)$ adalah Nilai gain ratio dari atribut A, yaitu ukuran yang digunakan untuk menentukan atribut terbaik sebagai pemisah (*split*) dalam algoritma C4.5. Gain ratio merupakan nilai Information Gain yang telah dinormalisasi. $IG(S,A)$ yaitu Information Gain dari atribut A terhadap himpunan data S, yaitu ukuran penurunan ketidakpastian (*entropy*) setelah data dipisahkan berdasarkan atribut A. $SplitInfo(A)$ adalah Nilai split information untuk atribut A, yaitu ukuran yang menggambarkan seberapa besar atribut A membagi data ke dalam beberapa subset. *SplitInfo* berfungsi sebagai faktor normalisasi untuk mengurangi bias Information Gain terhadap atribut dengan banyak nilai unik. A adalah Atribut yang sedang dievaluasi untuk ditentukan sebagai kandidat pemisah node.

2.5.2. Implementasi C4.5

Dalam penelitian ini, atribut yang dimasukkan ke dalam algoritma C4.5 berupa nilai numerik yang dihasilkan dari proses ekstraksi fitur TF-IDF pada setiap term dalam korpus[18]. Bobot TF-IDF tersebut direpresentasikan sebagai vektor fitur berdimensi besar, yang selanjutnya dianalisis oleh algoritma C4.5 untuk mengidentifikasi atribut yang paling berkontribusi dalam membedakan sentimen positif dan negatif[19]. Pemilihan atribut dilakukan secara bertahap dan berulang mulai dari node akar hingga terbentuk struktur pohon keputusan yang stabil, di mana setiap node daun (leaf node) berisi data yang sudah berada pada kelas yang sama atau homogen.

2.6 Evaluasi

Evaluasi dilakukan untuk mengetahui seberapa baik model Decision Tree C4.5 dalam mengklasifikasikan sentimen berdasarkan fitur TF-IDF[20]. Pada tahap ini, kinerja model diukur menggunakan beberapa metrik, yaitu presisi, recall, F1-score, dan akurasi untuk melihat performa secara keseluruhan. Selain itu, confusion matrix juga digunakan untuk melihat bagaimana model memprediksi setiap kelas, termasuk jumlah prediksi yang benar maupun yang salah. Proses

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 6, No 1, December 2025 | Hal 469-478
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v6i1.932

evaluasi akan dibagi menjadi rasio 70:30 dengan 70% data latih, 30% data uji. Selanjutnya 30% dibagi kembali dengan 15% data test dan 15% data validasi. Rumus yang digunakan pada bagian ini ditujukan oleh rumus persamaan 9,10,11,12.

a. Akurasi

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (9)$$

b. Presisi

$$Precision \frac{TP}{TP+FP} \quad (10)$$

c. Recall

$$Recall = \frac{TP}{TP+FN} \quad (11)$$

d. F1-Score

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision+Recall} \quad (12)$$

True Positive (TP) merupakan jumlah komentar positif yang berhasil diprediksi sebagai positif, sedangkan True Negative (TN) adalah jumlah komentar negatif yang berhasil diprediksi sebagai negatif. False Positive (FP) menunjukkan komentar negatif yang keliru diprediksi sebagai positif, sementara False Negative (FN) merepresentasikan komentar positif yang salah diprediksi sebagai negatif. Selain itu, Precision (P) digunakan untuk mengukur tingkat ketepatan prediksi positif yang dihasilkan sistem, Recall (R) mengukur kemampuan sistem dalam menemukan seluruh data positif yang sebenarnya, dan F1-Score (F1) merupakan nilai harmonis antara Precision dan Recall yang digunakan untuk mengevaluasi kinerja model klasifikasi secara keseluruhan.

3. HASIL DAN PEMBAHASAN

3.1 Dataset

Dataset yang digunakan pada penelitian ini terdiri dari 5.000 data terkait bitcoin yang diperoleh dari hasil webscraping di platform x dan telah diberi label positif dan negatif. Pada pelabelan di validasi oleh tiga anotator yaitu guru bahasa Indonesia. Pada keseluruhan data diperoleh 2.796 data positif dan 2.204 data negatif, data selanjutnya akan dibagi menjadi rasio 70:30 dengan 70% data latih, 30% data uji. Selanjutnya 30% dibagi kembali dengan 15% data test, 15% data validasi dan diperoleh hasil 3500 untuk data train, 750 data validasi dan 750 data test. Adapun Tampilan hasil labeling data bisa dilihat pada tabel 2.

Tabel 2 Tampilan hasil labeling data

No	tweet	label
1	Berita Bitcoin: 3 Alasan Kenapa BTC Masih Akan Tetap Turun di Bawah US$16.000 #bitcoin #kripto https://t.co/HU0DlFiV0a	negatif
2	@cryptondo Kapan bitcoin turun ke bawah lagi testing ke harga 10	negatif
...	...	...
4999	Institusi mulai lihatin Bitcoin lagi, vibe positif #144	positif
5000	BTC nge-dump tanpa peringatan, mental horor semua #2242 sih!!	negatif

Tabel 2 menampilkan contoh data komentar yang digunakan dalam penelitian ini, yang masing-masing telah diberi label sentimen positif maupun negatif. Setiap komentar merepresentasikan pendapat atau respons pengguna terhadap isu Bitcoin, dengan penentuan label dilakukan berdasarkan interpretasi isi teks: komentar yang bernuansa optimis atau mendukung dikategorikan sebagai positif, sedangkan komentar yang mengandung nada pesimis, kritik, atau kekhawatiran diklasifikasikan sebagai negatif. Data berlabel ini selanjutnya menjadi acuan dalam tahapan text preprocessing serta digunakan sebagai input utama pada proses pelatihan.

3.2 Text Preprocessing

Tahap preprocessing dilakukan untuk menyiapkan teks mentah sebelum masuk ke proses analisis. Proses ini mencakup Case folding, Penghapusan simbol, tanda baca dan emoji, Penghapusan URL, Penghapusan angka, Penghapusan karakter khusus, Tokenisasi, Stopword removal, Stemming. dapat dilihat pada tabel 3 sebelum dan sesudah preprocessing.

Tabel 3 Sebelum dan sesudah preprocessing

No	Sebelum di bersihkan	Sesudah dibersihkan
1	Berita Bitcoin: 3 Alasan Kenapa BTC Masih Akan Tetap Turun di Bawah US$16.000 #bitcoin #kripto https://t.co/HU0DlFiV0a	berita bitcoin alas btc tetap turun bawah us bitcoin kripto

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 6, No 1, December 2025 | Hal 469-478
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v6i1.932

No	Sebelum di bersihkan	Sesudah dibersihkan
2	@cryptondo Kapan bitcoin turun ke bawah lagi testing ke harga 10	cryptondo kapan bitcoin turun bawah testing harga
...	...	...
4999	Institusi mulai lihatin Bitcoin lagi, vibe positif #144	institusi mulai lihatin bitcoin vibe positif
5000	BTC nge-dump tanpa peringatan, mental horor semua #2242 sih!!	btc nge dump ingat mental horor semua sih

Tabel 3 menunjukkan perbedaan teks sebelum dan sesudah melalui proses text preprocessing. Pada bagian sebelum dibersihkan, teks masih berisi URL, mention, tagar, angka, dan berbagai simbol yang biasanya muncul di media sosial dan dapat mengganggu proses pengolahan data. Setelah dilakukan preprocessing seperti cleaning, case folding, dan penyeragaman kata, elemen-elemen tersebut dihilangkan sehingga teks menjadi lebih rapi dan fokus pada kata-kata yang penting. Dengan teks yang sudah bersih, proses ekstraksi fitur menggunakan TF-IDF dapat menangkap kata-kata yang benar-benar berpengaruh, sehingga hasilnya lebih akurat saat digunakan oleh algoritma Decision Tree C4.5. Perbandingan jumlah token dan kata unik sebelum dan sesudah preprocessing ditampilkan pada Tabel 4.

Tabel 4. Hasil visualisasi

Keterangan	Sebelum preprocessing	Sesudah preprocessing
Jumlah token	77847	58376
Jumlah kata unik	14852	4974

Hasil preprocessing menunjukkan adanya penurunan jumlah token dari 77847 menjadi 58376, serta penurunan jumlah kata unik dari 14852 menjadi 4974. Hal ini menunjukkan bahwa proses preprocessing berhasil mengurangi noise dan kata-kata yang tidak relevan pada seluruh dataset.

3.3 Hasil TF-IDF

Pada tahap ekstraksi fitur, seluruh data yang telah melalui proses preprocessing kemudian diubah ke dalam bentuk bobot menggunakan metode TF-IDF. Proses ini bertujuan untuk mengetahui seberapa penting suatu kata di dalam dokumen dibandingkan keseluruhan corpus. Kata dengan bobot TF-IDF tinggi dianggap lebih representatif dalam menggambarkan isi komentar. Tabel 5 berikut hanya menunjukkan sepuluh kata dengan nilai rata-rata TF-IDF berdasarkan hasil vektorisasi, yang nantinya digunakan sebagai fitur dalam proses klasifikasi sentimen.

Tabel 5. Hasil vektorisasi TF-IDF

No	Kata	Nilai TF-IDF
1	bitcoin	0.144607
2	btc	0.094382
3	naik	0.089786
4	turun	0.088491
5	harga	0.052765
6	mulai	0.047577
7	nih	0.037640
8	koreksi	0.035796
9	buat	0.034849
10	hari	0.032035

Berdasarkan hasil pada Tabel 5, terlihat bahwa kata “bitcoin” dan “btc” memiliki bobot TF-IDF yang paling tinggi, menandakan bahwa kata tersebut paling relevan dan paling berpengaruh dalam kumpulan data. Selain itu, kata seperti “naik”, “turun”, dan “harga” menunjukkan bahwa sebagian besar komentar berkaitan dengan kondisi pasar atau pergerakan harga Bitcoin. Bobot TF-IDF yang tinggi ini menjadi dasar kuat bagi model klasifikasi dalam membedakan sentimen positif dan negatif, karena kata-kata tersebut merepresentasikan konteks utama percakapan pengguna.
Selain melihat bobot TF-IDF rata-rata, penelitian ini juga menghitung total akumulasi nilai TF-IDF dari setiap kata di seluruh dokumen. Perhitungan ini digunakan untuk mengetahui kata mana yang paling dominan muncul dan paling sering memberikan kontribusi informasi pada dataset. Tabel 6 berikut menampilkan sepuluh kata total akumulasi bobot TF-IDF seluruh dokumen yang merepresentasikan kata paling sering digunakan dalam komentar terkait Bitcoin.

Tabel 6. Hasil akumulasi bobot TF-IDF

No	Kata	Total TF-IDF
1	bitcoin	723.033055
2	btc	471.910653
3	naik	448.931664
4	turun	442.454386

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 6, No 1, December 2025 | Hal 469-478
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v6i1.932

No	Kata	Total TF-IDF
5	harga	263.823078
6	mulai	237.883089
7	nih	188.198393
8	koreksi	178.978814
9	buat	174.244238
10	hari	160.173993

Hasil pada Tabel 6 menunjukkan bahwa kata "bitcoin" dan "btc" kembali menempati posisi teratas dengan total skor TF-IDF tertinggi, yang menunjukkan frekuensi kemunculan yang tinggi sekaligus relevansi yang kuat di dalam dataset. Kata seperti "naik", "turun", dan "harga" juga tampil konsisten sebagai kata dominan, menandakan bahwa para pengguna banyak membahas kenaikan atau penurunan harga Bitcoin. Temuan ini memberikan gambaran umum mengenai pola bahasa dan fokus utama dalam percakapan pengguna, serta mendukung proses klasifikasi karena kata-kata dominan ini berperan besar dalam pembentukan struktur pohon keputusan menggunakan algoritma C4.5. Tabel 5 menampilkan nilai rata-rata bobot TF-IDF, sedangkan Tabel 6 menampilkan total akumulasi bobot TF-IDF seluruh dokumen

3.4 Hasil Klasifikasi Decision Tree C45

Tabel berikut menyajikan hasil pembentukan struktur *decision tree C4.5* berdasarkan data *train* yang telah melalui proses preprocessing dan ekstraksi fitur TF-IDF. Setiap node menggambarkan atribut (fitur kata) yang dipilih sebagai pemisah berdasarkan nilai Gain Ratio tertinggi pada tahap tersebut. Node dengan tipe leaf menunjukkan kelas akhir (positif atau negatif) yang ditetapkan oleh model berdasarkan mayoritas data pada cabang tersebut. Informasi yang dicantumkan meliputi ID node, ID parent, nilai cabang, jenis node, fitur pemisah, nilai Gain Ratio, kelas prediksi, serta jumlah data pada node tersebut. Penyajian ini memberikan gambaran yang lebih terstruktur mengenai alur pengambilan keputusan oleh algoritma C4.5.

Tabel 7. Hasil konversi decision tree c45

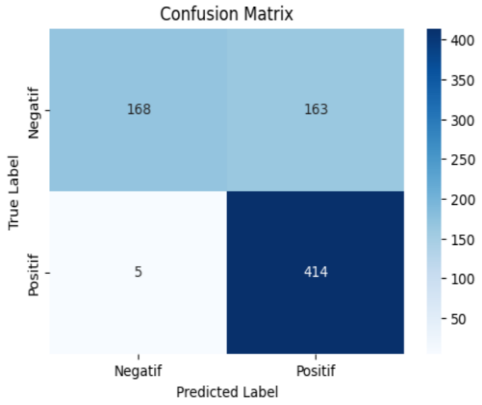
No	Node ID	Parent ID	Branch Value	Node Type	Feature	Gain Ratio	Class	Jumlah Data (n)
1	1	NaN	NaN	Node	spike	0.1911	-	3500
2	2	1.0	0.0	Node	tekan	0.2048	-	3385
3	3	2.0	0.0	Node	keok	0.2239	-	3197
4	4	3.0	0.0	Node	rawan	0.2336	-	3086
5	5	4.0	0.0	Node	temu	0.2473	-	2986
6	6	5.0	0.0	Node	nge	0.2605	-	2890
7	7	6.0	0.0	Leaf	-	-	1	2800
8	8	6.0	1.0	Leaf	-	-	0	90
9	9	5.0	1.0	Leaf	-	-	0	96
10	10	4.0	1.0	Leaf	-	-	0	100

Berdasarkan tampilan struktur pohon pada tabel, terlihat bahwa model C4.5 memilih fitur-fitur tertentu seperti *spike*, *tekan*, *keok*, *rawan*, *temu*, dan *nge* sebagai pemisah utama dalam proses klasifikasi. Pemilihan fitur tersebut menunjukkan bahwa kata-kata tersebut memiliki pengaruh paling besar dalam membedakan komentar positif dan negatif pada dataset Bitcoin. Node *leaf* yang muncul pada tingkat-tingkat akhir menunjukkan hasil pengelompokan yang sudah tidak dapat dipecah lagi karena semua datanya masuk ke kelas yang sama. Pada tabel tersebut, nilai class 1 merepresentasikan sentimen positif, sedangkan nilai class 0 merepresentasikan sentimen negatif. Simbol (-) pada kolom class menunjukkan bahwa node tersebut merupakan node internal yang belum menghasilkan keputusan kelas akhir. Struktur pohon ini kemudian digunakan oleh model untuk melakukan prediksi sentimen pada data uji secara lebih sistematis dan dapat ditelusuri kembali proses pengambilannya.

3.5 Evaluasi

Data latih digunakan untuk membangun model Decision Tree C4.5. Data validasi digunakan untuk mengevaluasi performa model selama proses pelatihan dan membantu mencegah overfitting, sedangkan data uji digunakan untuk mengukur kinerja akhir model terhadap data yang benar-benar baru. Pada proses evaluasi model, data penelitian dibagi menggunakan skema rasio awal 70:30, di mana 70% data train, sedangkan 30% sisanya diproses lebih lanjut menjadi data validasi dan data test. Dari 30% tersebut, dibagi kembali menjadi 15% untuk validasi dan 15% untuk test. Pembagian seperti ini bertujuan agar model C4.5 bisa dilatih secara optimal, sekaligus diuji kestabilannya menggunakan data validasi dan data uji yang belum pernah dilihat model sebelumnya. Setelah proses pelatihan selesai, performa model dianalisis menggunakan Confusion Matrix, yang berfungsi untuk melihat jumlah prediksi benar dan salah pada masing-masing kelas sentimen. Visualisasi ini penting karena memberikan gambaran yang lebih jelas mengenai akurasi serta kemampuan model dalam mengenali perbedaan antara sentimen positif dan negatif. Adapun gambar 1 menampilkan confusion matrix.

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 6, No 1, December 2025 | Hal 469-478
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v6i1.932



Gambar 1. Confusion matrix decision tree c45(70:30)

Berdasarkan Gambar 1, dapat dilihat bahwa model C4.5 memiliki performa klasifikasi yang cukup baik. Model mampu mengidentifikasi kelas positif dengan akurasi tinggi, ditunjukkan oleh jumlah prediksi positif yang benar jauh lebih besar dibandingkan prediksi positif yang salah. Sementara itu, pada kelas negatif masih ditemukan jumlah kesalahan prediksi yang lebih tinggi, sehingga menunjukkan bahwa model sedikit lebih sensitif terhadap kelas positif. Secara keseluruhan, hasil evaluasi dari data uji (yang merupakan 15% dari dataset) menunjukkan bahwa model Decision Tree C4.5 bekerja cukup efektif dalam mengklasifikasikan sentimen pada data tweet. Meskipun demikian, masih ada ruang untuk meningkatkan kinerja model, terutama dalam menyeimbangkan kemampuan prediksi antara kelas positif dan negatif. Adapun hasil yang didapat ini akan ditampilkan pada tabel test evaluasi.

Tabel berikut menyajikan hasil evaluasi model Decision Tree C4.5 pada data komentar yang telah dibagi menggunakan rasio pelatihan dan pengujian sebesar 70:30. Evaluasi dilakukan menggunakan metrik Precision, Recall, dan F1-Score untuk masing-masing kelas, yaitu negatif dan positif. Selain itu, ditampilkan juga nilai accuracy, serta rata-rata macro dan weighted untuk memberikan gambaran menyeluruh terhadap performa model. Nilai-nilai ini digunakan untuk mengetahui seberapa baik model mampu melakukan klasifikasi sentimen pada data uji, khususnya dalam membedakan komentar negatif dan positif.

Tabel 8. Hasil test evaluasi decision tree c45

Rasio		Precision	Recall	F1-Score	Support
	negatif	1.00	0.49	0.66	110
	positif	0.71	1.00	0.83	140
90:10	accuracy			0.78	250
	macro avg	0.86	0.75	0.75	250
	Weighted avg	0.84	0.78	0.76	250
	negatif	1.00	0.50	0.66	221
	positif	0.72	1.00	0.83	279
80:20	accuracy			0.78	500
	macro avg	0.86	0.75	0.75	500
	Weighted avg	0.84	0.78	0.76	500
	negatif	0.97	0.51	0.67	331
	positif	0.72	0.99	0.83	419
70:30	accuracy			0.78	750
	macro avg	0.84	0.75	0.75	750
	Weighted avg	0.83	0.78	0.76	750

Berdasarkan hasil evaluasi pada Tabel 8, terlihat adanya perbedaan yang cukup signifikan antara nilai recall pada kelas positif (0.99) dan kelas negatif (0.51). Kondisi ini mengindikasikan bahwa model Decision Tree C4.5 cenderung bias terhadap kelas positif, di mana sebagian besar data uji diprediksi sebagai sentimen positif. Bias ini dipengaruhi oleh distribusi data yang tidak sepenuhnya seimbang, dengan jumlah data positif yang lebih dominan dibandingkan data negatif, sehingga model lebih sensitif terhadap pola pada kelas mayoritas. Selain itu, nilai precision yang sangat tinggi pada kelas negatif (hingga 0.97–1.00) menunjukkan bahwa model sangat berhati-hati dalam memprediksi sentimen negatif dan hanya memberikan prediksi negatif pada data yang memiliki karakteristik yang sangat kuat. Meskipun nilai precision yang tinggi menunjukkan tingkat kesalahan yang rendah pada prediksi negatif, kondisi ini juga menyebabkan rendahnya nilai recall pada kelas tersebut. Fenomena ini mencerminkan adanya trade-off antara precision dan recall, yang umum terjadi pada kasus klasifikasi data teks media sosial yang tidak seimbang. Oleh karena itu, hasil ini menunjukkan bahwa meskipun model memiliki tingkat kepercayaan tinggi terhadap prediksi negatif, kemampuannya dalam mendeteksi seluruh data negatif masih terbatas. Penelitian selanjutnya dapat mempertimbangkan penerapan teknik penanganan data tidak seimbang atau metode regularisasi untuk mengurangi bias model dan meningkatkan keseimbangan performa antar kelas.

4. KESIMPULAN

Penelitian ini menunjukkan bahwa algoritma Decision Tree C4.5 mampu mengklasifikasikan sentimen komentar pengguna terkait Bitcoin dengan performa yang cukup baik. Setelah melalui tahapan preprocessing dan ekstraksi fitur menggunakan TF-IDF, model menghasilkan tingkat akurasi sebesar 78% pada data uji. Hasil evaluasi per kelas menunjukkan bahwa model sangat baik dalam mengenali komentar positif dengan nilai recall sebesar 0.99, namun masih kurang optimal dalam mendeteksi komentar negatif yang hanya mencapai recall sebesar 0.51, meskipun nilai precision pada kelas negatif tergolong sangat tinggi, yaitu sebesar 0.97. Rendahnya nilai recall pada kelas negatif disinyalir dipengaruhi oleh ketidakseimbangan distribusi data (data imbalance) pada dataset, di mana jumlah data positif lebih dominan dibandingkan data negatif, sehingga model cenderung lebih sensitif terhadap kelas mayoritas. Dengan demikian, tujuan penelitian untuk menganalisis sentimen masyarakat terhadap Bitcoin melalui klasifikasi komentar media sosial telah tercapai. Penelitian selanjutnya dapat mempertimbangkan penggunaan algoritma pohon keputusan yang lebih kompleks atau penerapan teknik ekstraksi fitur yang lebih kaya untuk meningkatkan kualitas dan keseimbangan hasil prediksi..

REFERENCES

[1] J. E. Savero, V. H. Pranatawijaya, and E. Christian, "Analisis Sentimen Pengguna Media Sosial X terhadap Perubahan Harga Bitcoin: Pendekatan Machine Learning," *Konvergensi Teknologi dan Sistem Informasi*, vol. 4, no. 1, 2024, doi: 10.24002/konstelasi.v4i1.9043.

[2] I. Septiningsih, "Analisis Cryptocurrency sebagai Instrumen Transaksi di Indonesia," *Legal Advice Journal Of Law*, vol. 1, no. 2, 2024, doi: 10.12345

[3] M. H. Widianto and Y. Cornelius, "Sentiment Analysis towards Cryptocurrency and NFT in Bahasa Indonesia for Twitter Large Amount Data Using BERT," *Original Research Paper International Journal of Intelligent Systems and Applications in Engineering IJISAE*, vol. 2023, no. 1, pp. 303–309, 2023, Available: https://orcid.org/0000-0001-8722-9868

[4] R. Syahputra and Y. M. Putra, "Implementation Of The C4.5 Algorithm In Describing The Trends Of The Human Consciousness And Unconsciousness," *Journal of Applied Engineering and Technological Science*, vol. 3, no. 2, pp. 208–213, 2022, doi : 10.37385/jaets.v3i2.841.

[5] A. Sentimen *et al.*, "Sentiment Analysis of Cryptocurrency Exchange Application on Twitter Using Naïve Bayes Classifier Method," *Jurnal Informatika dan Teknologi Informasi*, vol. 20, no. 1, pp. 15–30, 2023, doi: 10.31515/telematika.v20i1.9044.

[6] M. Solehuddin, W. A. Syafei, and R. Gernowo, "Metode Decision Tree untuk Meningkatkan Kualitas Rencana Pelaksanaan Pembelajaran dengan Algoritma C4.5," *Jurnal Penelitian dan Pengembangan Pendidikan*, vol. 6, no. 3, pp. 510–519, Oct. 2022, doi: 10.23887/jppp.v6i3.52840.

[7] I. T. Julianto, D. Kurniadi, M. R. Nashrulloh, and A. Mulyani, "TWITTER SOCIAL MEDIA SENTIMENT ANALYSIS AGAINST BITCOIN CRYPTOCURRENCY TRENDS USING RAPIDMINER," *Jurnal Teknik Informatika (Jutif)*, vol. 3, no. 5, pp. 1183–1187, Oct. 2022, doi: 10.20884/1.jutif.2022.3.5.289.

[8] I. D. Mienye and N. Jere, "A Survey of Decision Trees: Concepts, Algorithms, and Applications," *IEEE Access*, vol. 12, pp. 86716–86727, 2024, doi: 10.1109/ACCESS.2024.3416838.

[9] I. Habib Kusuma and N. Cahyono, "Analisis Sentimen Masyarakat Terhadap Penggunaan E-Commerce Menggunakan Algoritma K-Nearest Neighbor," *Jurnal Informatika Jurnal Pengembangan IT*, vol. 8, no. 3, pp. 302-307, 2023, doi : 10.30591/jpit.v8i3.5734.

[10] R. Merdiansah and A. Ali Ridha, "Analisis Sentimen Pengguna X Indonesia Terkait Kendaraan Listrik Menggunakan IndoBERT," *Jurnal Ilmu Komputer dan Sistem Informasi (JIKOMSI*, vol. 7, no. 1, pp. 221–228, 2024, doi: 10.55338/jikomsi.v7i1.2895.

[11] P. Arsi and R. Waluyo, "Analisis Sentimen Wacana Pemindahan Ibu Kota Indonesia Menggunakan Algoritma Support Vector Machine (SVM)," *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 8, no. 1, pp. 147–156, 2021, doi: 10.25126/jtiik.202183944.

[12] D. E. Cahyani and I. Patasik, "Performance comparison of tf-idf and word2vec models for emotion text classification," *Bulletin of Electrical Engineering and Informatics*, vol. 10, no. 5, pp. 2780–2788, Oct. 2021, doi: 10.11591/eei.v10i5.3157.

[13] H. Zhou, "Research of Text Classification Based on TF-IDF and CNN-LSTM," *J Phys Conf Ser*, vol. 2171, no. 1, Jan. 2022, doi: 10.1088/1742-6596/2171/1/012021.

[14] F. N. H. Alfandi Safira, "Analisis Sentimen Masyarakat Terhadap Paylater Menggunakan Metode Naive Bayes Classifier," *Jurnal Sistem Informasi*, vol. 5, no. 1, 2023, doi: 10.31849/zn.v5i1.12856.

[15] A. N. Halim, R. Rudiman, and N. A. Verdikha, "Analisis Sentimen Opini Publik Terhadap Peristiwa Bitcoin Halving Pada Data Teks Twitter Menggunakan Metode Naïve Bayes Dan Pembobotan Fitur TF-IDF," *RIGGS: Journal of Artificial Intelligence and Digital Business*, vol. 4, no. 3, pp. 2823–2831, Aug. 2025, doi: 10.31004/riggs.v4i3.2291.

[16] T. A. Assegie, R. L. Tulasi, and N. K. Kumar, "Breast cancer prediction model with decision tree and adaptive boosting," *IAES International Journal of Artificial Intelligence*, vol. 10, no. 1, pp. 184–190, 2021, doi: 10.11591/ijai.v10.i1.pp184-190.

[17] T. Johnson and S. Shamroukh, "Predictive modeling of burnout based on organizational culture perceptions among health systems employees: a comparative study using correlation, decision tree, and Bayesian analyses," *Sci Rep*, vol. 14, no. 1, Dec. 2024, doi: 10.1038/s41598-024-56771-2.

[18] Y Partogi, A Pasaribu, Sutrisno., "Perancangan Metode Decision Tree Terhadap Sistem Perpustakaan Stmik Kuwera," *Jurnal Sistem Informasi dan Teknologi (SINTEK)*, vol. 1, no. 2, p. 20, 2021, doi: 10.56995/sintek.v1i2.4.

[19] O. Barukab, A. Ahmad, T. Khan, and M. R. Thayyil Kunhumuhammed, "Analysis of Parkinson's Disease Using an Imbalanced-Speech Dataset by Employing Decision Tree Ensemble Methods," *Diagnostics*, vol. 12, no. 12, Dec. 2022, doi: 10.3390/diagnostics12123000.

[20] S. A. Arnomo, A. A. Fajrin, Y. Siyamto, S. Fairuz, and N. Sadikin, "Evaluasi Model Decision Tree Pada Keputusan Kelayakan Kredit," *Jurnal Desain Dan Analisis Teknologi (JDDAT)*, vol. 2, no. 2, 2023, doi: 10.58520/jddat.v2i2.39.