ANALISIS SENTIMEN MASYARAKAT TERHADAP PENGGUNDULAN HUTAN DAN BENCANA ALAM PADA PLATFORM MEDIA SOSIAL MENGGUNAKAN ALGORITMA C4.5

Putra Frianto[1], Bayu Rianto[2], Usman[3]
[1,2,3] Program Studi Sistem Informasi, Fakultas Teknik dan Ilmu Komputer, Universitas Islam Indragiri
JL. Provinsi Parit 1 Tembilahan Hulu, Indragiri Hilir, Riau, Indonesia
Email: [1]putrafrianto0906@gmail.com, [2]rianto.bayu91@gmail.com, [3]usmanovsky13411@gmail.com

ABSTRAK

Media sosial khususnya platform X (dahulu Twitter), kini menjadi instrumen krusial bagi publik dalam menyuarakan opini maupun respons terhadap fenomena lingkungan secara masif. Penelitian ini dilatarbelakangi oleh urgensi analisis sentimen terkait isu deforestasi dan bencana alam, yang dipicu oleh meluasnya dampak ekologis, kerugian materil, serta dinamika kritik terhadap kebijakan mitigasi pemerintah. Melalui identifikasi pandangan publik, pemangku kebijakan dapat memetakan kegelisahan masyarakat sekaligus menyusun strategi penanggulangan lingkungan yang lebih akomodatif. Studi ini bertujuan mengimplementasikan Algoritma C4.5 untuk mengklasifikasikan kecenderungan sentimen masyarakat mengenai topik tersebut menjadi beberapa klaster opini. Proses diawali melalui tahapan preprocessing data yang meliputi eliminasi tautan, penghapusan karakter non-alfanumerik, pembersihan stopwords, dan tokenisasi. Ekstraksi fitur teks kemudian dioptimalkan menggunakan pembobotan Term Frequency-Inverse Document Frequency (TF-IDF) untuk mengonversi korpus menjadi matriks numerik. Evaluasi performa diuji melalui dua skenario klasifikasi, yakni kategorisasi biner (positif dan negatif) serta multi-kelas (positif, negatif, dan netral). Temuan eksperimen menunjukkan bahwa pemodelan dengan dua kelas klasifikasi diperoleh accuracy sebesar 99,17%, precision sebesar 99,17%, recall sebesar 99,17%, dan F1-score sebesar 99,17%. Sementara itu, pada klasifikasi tiga kelas diperoleh accuracy sebesar 98,76%, precision sebesar 98,76%, recall sebesar 99,17%, dan F1-score sebesar 98,96%. Riset ini menyimpulkan bahwa algoritma C4.5 dapat memberikan akurasi yang cukup valid ketika diterapkan pada model klasifikasi sentimen biner dalam menganalisis diskursus lingkungan di media sosial.

Kata Kunci : Analisis Sentimen, Deforestasi, Bencana Alam, Algoritma C4.5, TF-IDF.

1 PENDAHULUAN

Perkembangan teknologi informasi telah mengubah pola komunikasi masyarakat, termasuk dalam penyampaian opini terhadap berbagai isu sosial dan lingkungan. Media sosial menjadi salah satu sumber data yang dapat digunakan untuk melihat bagaimana masyarakat membicarakan suatu isu secara terbuka. Twitter, yang kini dikenal sebagai X, telah banyak dimanfaatkan dalam penelitian untuk mengkaji opini dan diskursus publik karena menyediakan sejumlah besar unggahan yang dapat dianalisis berdasarkan topik tertentu. Penelitian mengenai isu lingkungan juga menunjukkan bahwa data dari Twitter/X dapat digunakan untuk mengidentifikasi pola pembahasan dan kecenderungan sentimen masyarakat terhadap persoalan lingkungan [1]

Salah satu isu lingkungan yang banyak mendapat perhatian adalah penggundulan hutan (deforestasi) dan keterkaitannya dengan bencana alam. Deforestasi dapat memengaruhi kondisi lingkungan dan sistem hidrologi suatu wilayah, sehingga dalam kondisi tertentu dapat berkaitan dengan terjadinya bencana seperti banjir. Penelitian mengenai hubungan deforestasi dan kejadian banjir di Daerah Aliran Sungai Lasolo menunjukkan adanya pola keterkaitan antara perubahan tutupan hutan dan kejadian banjir pada wilayah tersebut [2]. Di sisi lain, isu lingkungan seperti perubahan penggunaan lahan dan hilangnya keanekaragaman hayati juga menjadi bagian dari

diskursus masyarakat di media sosial. Hal tersebut menunjukkan bahwa media sosial dapat menjadi sumber informasi untuk mengamati perhatian dan pandangan masyarakat terhadap persoalan lingkungan.

Banyaknya opini masyarakat yang berbentuk teks dan tidak terstruktur menyebabkan diperlukan metode komputasi untuk mengolah data tersebut secara sistematis. Analisis sentimen merupakan salah satu pendekatan yang dapat digunakan untuk mengidentifikasi kecenderungan opini dalam data teks ke dalam kategori sentimen tertentu. Penelitian mengenai sentimen terhadap isu keberlanjutan menunjukkan bahwa data media sosial dapat diolah untuk memperoleh gambaran mengenai sikap atau kecenderungan masyarakat terhadap isu lingkungan. Selain itu, penelitian terhadap sentimen masyarakat Indonesia mengenai kebijakan keanekaragaman hayati juga menunjukkan bahwa Twitter dapat digunakan sebagai sumber data untuk mengkaji persepsi masyarakat terhadap isu lingkungan [3][4].

Dalam penelitian ini, analisis sentimen dilakukan menggunakan algoritma C4.5 sebagai metode klasifikasi berbasis Decision Tree. Algoritma C4.5 menggunakan proses pemilihan atribut berdasarkan information gain untuk membentuk aturan klasifikasi. Pada penelitian analisis sentimen, algoritma Decision Tree C4.5 juga telah digunakan untuk mengklasifikasikan data teks dari media sosial. Hidayat et al, misalnya, menerapkan Decision Tree C4.5 pada analisis sentimen di Twitter dengan data yang diberi label positif dan negatif. Sebelum proses klasifikasi, data teks dapat diubah menjadi bentuk numerik menggunakan pembobotan Term Frequency-Inverse Document Frequency (TF-IDF). Penelitian mengenai analisis sentimen isu lingkungan pada platform X juga menunjukkan penggunaan tahapan text cleaning, tokenization, stopword removal, stemming, dan TF-IDF sebelum proses klasifikasi [5].

Berdasarkan uraian tersebut, penelitian ini bertujuan untuk menganalisis sentimen masyarakat terhadap isu penggundulan hutan dan bencana alam pada media sosial menggunakan algoritma C4.5. Penelitian dilakukan melalui tahapan preprocessing, pembobotan TF-IDF, dan klasifikasi menggunakan C4.5. Pengujian dilakukan dengan dua skenario klasifikasi, yaitu dua kelas sentimen yang terdiri dari positif dan negatif serta tiga kelas sentimen yang terdiri dari positif, negatif, dan netral. Hasil penelitian diharapkan dapat memberikan gambaran mengenai kecenderungan sentimen masyarakat terhadap isu penggundulan hutan dan bencana alam berdasarkan data media sosial yang dianalisis.

2 TINJAUAN PUSTAKA

2.1 Analisis Sentimen

Analisis sentimen merupakan proses untuk mengidentifikasi dan mengklasifikasikan opini dalam data teks ke dalam kategori tertentu, seperti positif, negatif, dan netral. Analisis sentimen dapat digunakan untuk mengetahui kecenderungan opini masyarakat terhadap suatu topik berdasarkan data media sosial [6].

Pada penelitian ini, analisis sentimen digunakan untuk mengklasifikasikan opini masyarakat mengenai penggundulan hutan dan bencana alam ke dalam dua skenario, yaitu kelas positif-negatif dan kelas positif-negatif-netral.

2.2 Text Preprocessing

Text preprocessing dilakukan untuk membersihkan dan menyeragamkan data teks sebelum proses klasifikasi. Tahapan yang digunakan meliputi:

- Cleaning, yaitu menghapus tautan, simbol, dan karakter yang tidak diperlukan.
- Case folding, yaitu mengubah seluruh huruf menjadi huruf kecil.
- Tokenization, yaitu memecah teks menjadi kata atau token.
- Stopword removal, yaitu menghapus kata umum yang kurang memberikan informasi.
- Stemming, yaitu mengubah kata berimbuhan menjadi bentuk dasarnya, seperti "penebangan" menjadi "tebang".

Tahapan tersebut dilakukan untuk menghasilkan data teks yang lebih terstruktur dan siap digunakan pada proses ekstraksi fitur. Definisi stemming ini juga memperbaiki kekeliruan pada naskah sebelumnya yang masih mencampurkannya dengan stopword removal.

2.3 TF-IDF

Term Frequency-Inverse Document Frequency (TF-IDF) merupakan metode pembobotan yang mengubah data teks menjadi bentuk numerik berdasarkan frekuensi kemunculan kata dalam dokumen dan keseluruhan dokumen. Nilai TF-IDF dihitung dengan:

$$TF{-}IDF(t,d)=TF(t,d)\times IDF(t)$$

dengan:

$$IDF(t)=\log\left(\frac{N}{dft}\right)$$

Keterangan: N adalah jumlah dokumen dan dft adalah jumlah dokumen yang mengandung term t. Hasil pembobotan TF-IDF digunakan sebagai input dalam proses klasifikasi.

2.4 Algoritma C4.5

C4.5 merupakan algoritma klasifikasi berbasis Decision Tree yang dikembangkan dari ID3. Algoritma ini membentuk pohon keputusan berdasarkan pemilihan atribut dan menggunakan gain ratio dalam menentukan atribut pemisah. C4.5 juga memiliki mekanisme pruning untuk mengurangi kompleksitas pohon [7].

Perhitungan pada C4.5 meliputi entropy, information gain, dan gain ratio. Rumus entropy adalah:

$$Entropy(S)=-\sum_{i=1}^{n} log_2\ (p_i)$$

Sedangkan gain ratio dihitung berdasarkan perbandingan information gain dengan split information. Dalam penelitian ini, C4.5 digunakan untuk mengklasifikasikan data sentimen berdasarkan fitur yang telah diperoleh melalui TF-IDF.

Penelitian ini tidak menyatakan C4.5 lebih baik atau lebih akurat dibandingkan algoritma lain karena tidak dilakukan pengujian dengan algoritma pembanding.

2.5 Confusion Matrix

Confusion matrix digunakan untuk mengevaluasi hasil klasifikasi dengan membandingkan kelas aktual dan hasil prediksi. Metrik yang digunakan meliputi accuracy, precision, recall, dan F1-score.

Rumus accuracy yang digunakan adalah:

$$Accuracy=\frac{TP+TN}{TP+TN+FP+FN}X100\%$$

Penggunaan rumus ini sekaligus memperbaiki kesalahan pada naskah sebelumnya yang menuliskan "TN + TN + TP".

2.6 Ketidakseimbangan Kelas

Class imbalance terjadi ketika jumlah data pada setiap kelas tidak seimbang. Kondisi ini dapat menyebabkan accuracy menjadi kurang menggambarkan kemampuan model dalam mengenali kelas minoritas. Oleh karena itu, evaluasi perlu memperhatikan precision, recall, dan F1-score pada masing-masing kelas. Pada penelitian ini, data hasil pelabelan terdiri dari 1.595 data negatif, 15 data positif, dan 7 data netral dari total 1.617 data. Kondisi tersebut menunjukkan adanya ketidakseimbangan kelas yang perlu dipertimbangkan dalam interpretasi hasil klasifikasi.

3 METODE PENELITIAN

Penelitian ini menggunakan pendekatan kuantitatif dengan metode eksperimen komputasi. Penelitian dilakukan untuk menganalisis sentimen masyarakat terhadap isu penggundulan hutan dan bencana alam pada platform media sosial menggunakan algoritma C4.5. Data yang digunakan

berupa data teks yang diperoleh dari media sosial dan berkaitan dengan kata kunci serta hashtag mengenai penggundulan hutan dan bencana alam.

Tahapan penelitian terdiri dari pengumpulan data, text preprocessing, pelabelan sentimen, pembobotan TF-IDF, pembagian data menjadi data latih dan data uji, proses klasifikasi menggunakan algoritma C4.5, serta evaluasi menggunakan confusion matrix. Pada penelitian ini dilakukan dua skenario klasifikasi, yaitu klasifikasi 2 kelas yang terdiri dari positif dan negatif, serta klasifikasi 3 kelas yang terdiri dari positif, negatif, dan netral. Alur tersebut digunakan untuk mengetahui kemampuan algoritma C4.5 dalam mengklasifikasikan opini masyarakat berdasarkan data teks yang diperoleh.

3.1 Sumber Data

Data dalam penelitian ini bersumber dari media sosial yang digunakan masyarakat untuk menyampaikan pendapat mengenai isu penggundulan hutan dan bencana alam. Pengumpulan data dilakukan melalui proses crawling dengan menggunakan kata kunci dan hashtag yang berhubungan dengan topik penelitian. Data awal yang diperoleh berjumlah 2.363 data. Selanjutnya data tersebut melalui proses preprocessing untuk menghilangkan data yang tidak diperlukan. Setelah proses tersebut diperoleh 1.619 data yang layak digunakan. Setelah dilakukan pemeriksaan dan pelabelan, diperoleh 1.617 data akhir yang digunakan dalam proses analisis sentimen.

Data akhir tersebut kemudian dikelompokkan berdasarkan sentimen menjadi tiga kategori, yaitu positif, negatif, dan netral. Distribusi data yang digunakan adalah sebagai berikut.

No	Sentimen	Jumlah Data	Persentase
1	Negatif	1.595	98,64%
2	Positif	15	0,93%
3	Netral	7	0,43%
	Total	1.617	100%

Persentase masing-masing kelas dihitung dengan rumus:

$$\text{Persentase} = \frac{Jumlah\ Data}{Total\ Data} \times 100\%$$

Untuk sentimen negatif:

$$\frac{1595}{1617} \times 100\% = 98{,}64\%$$

Untuk sentimen positif:

$$\frac{15}{1617} \times 100\% = 0{,}93\%$$

Untuk sentimen netral:

$$\frac{7}{1617} \times 100\% = 0{,}43\%$$

Dengan demikian, jumlah keseluruhan data adalah:

$$1595 + 15 + 7 = 1617$$

Data menunjukan bahwa sentiment negatif memiliki jumlah paling besar dibandingkan sentimen positif dan netral.

3.2 Text Preprocessing

Sebelum digunakan dalam proses klasifikasi, data teks terlebih dahulu dilakukan text preprocessing. Tahapan ini bertujuan untuk membersihkan data sehingga teks yang digunakan dalam proses analisis menjadi lebih terstruktur.

3.3 Pelabelan Sentimen

Setelah data melalui tahap preprocessing, data diberi label berdasarkan kecenderungan sentimennya. Pada penelitian ini digunakan dua skenario pelabelan. Skenario pertama menggunakan 2 kelas, yaitu:

- Positif
- Negatif

Skenario kedua menggunakan 3 kelas, yaitu:

- Positif
- Negatif
- Netral

Penggunaan dua skenario tersebut bertujuan untuk mengetahui perbedaan hasil klasifikasi apabila data netral tidak digunakan dan apabila data netral tetap dipertahankan dalam proses klasifikasi.

3.4 Pembobotan TF-IDF

Tahap pertama adalah menghitung Term Frequency (TF), yaitu frekuensi kemunculan suatu kata dalam sebuah dokumen. Selanjutnya dihitung nilai Inverse Document Frequency (IDF) untuk mengetahui tingkat kepentingan kata dalam keseluruhan dokumen.

Rumus IDF adalah:

$$\text{IDF(t)}=\log\left(\frac{N}{dft}\right)$$

Keterangan:

- N = jumlah keseluruhan dokumen
- df(t) = jumlah dokumen yang mengandung kata tt

Selanjutnya nilai TF-IDF dihitung menggunakan:

$$\text{TF–IDF(t,d)}=\text{TF(t,d)}\times\text{IDF(t)}$$

Keterangan:

- TF(t,d) = frekuensi kata t pada dokumen d
- IDF(t) = nilai Inverse Document Frequency
- TF-IDF(t,d) = bobot kata t pada dokumen d

Hasil pembobotan tersebut digunakan sebagai data masukan pada proses klasifikasi menggunakan algoritma C4.5. Naskah Putra juga menggunakan pembobotan TF-IDF sebelum proses klasifikasi.

3.5 Algoritma C4.5

Algoritma C4.5 digunakan untuk melakukan klasifikasi terhadap data sentimen yang telah melalui tahap preprocessing dan pembobotan TF-IDF. Algoritma ini membentuk pohon keputusan berdasarkan atribut yang memiliki nilai informasi terbaik.

Tahapan C4.5 dimulai dengan menghitung Entropy. Rumus Entropy adalah:

$$\text{Entropy(S)}=-\sum_{i=1}^{n} log_2\,(p_i)$$

Keterangan:

- S = himpunan data
- pi = proporsi data pada kelas ke-i
- n = jumlah kelas

Setelah nilai Entropy diperoleh, selanjutnya dilakukan perhitungan Gain.

$$\text{Gain(S,A)}=\text{Entropy(S)} - \sum_{v\in Values(A)} \frac{|Sv|}{|S|}\,\text{Entropy}\,(sv)$$

Keterangan:

- S = keseluruhan data

- A = atribut
- Sv = subset data berdasarkan atribut AA
- |S| = jumlah keseluruhan data
- |Sv| = jumlah data pada subset

Atribut yang mempunyai nilai Gain paling tinggi digunakan sebagai atribut pemisah pada pohon keputusan. Proses tersebut dilakukan sampai terbentuk aturan klasifikasi yang dapat digunakan untuk menentukan kelas sentimen data.

3.6 Pembagian Data

Data sebanyak 1.617 kemudian dibagi menjadi data latih dan data uji. Penelitian menggunakan tiga skenario pembagian data, yaitu 90%:10%, 80%:20%, dan 70%:30%.

3.6.1 Pembagian 90% : 10%

Jumlah data latih:

$$1617 \times 90\% = 1455{,}3 \approx 1455$$

Jumlah data uji:

$$1617 \times 10\% = 161{,}7 \approx 162$$

Sehingga:

Jenis Data	Jumlah
Data latih	1.455
Data uji	162
Total	1.617

3.6.2 Pembagian 80% : 20%

Jumlah data latih:

$$1617 \times 80\% = 1293{,}6 \approx 1293$$

Jumlah data uji:

$$1617 \times 20\% = 323{,}4 \approx 324$$

Jenis Data	Jumlah
Data latih	1.293
Data uji	324
Total	1.617

3.6.3 Pembagian 70% : 30%

Jumlah data latih:

$$1617 \times 70\% = 1131{,}9 \approx 1132$$

Jumlah data uji:

$$1617 \times 30\% = 485{,}1 \approx 485$$

Jenis Data	Jumlah
Data latih	1.132
Data uji	485
Total	1.617

Dengan demikian, skema pembagian data penelitian adalah:

Pembagian	Data Latih	Data Uji
90% : 10%	1.455	162
80% : 20%	1.293	324

Pembagian	Data Latih	Data Uji
70% : 30%	1.132	485

3.7 Confusion Matrix

Setelah proses klasifikasi menggunakan C4.5 selesai, hasil prediksi dibandingkan dengan kelas aktual menggunakan confusion matrix. Confusion matrix digunakan untuk mengetahui jumlah prediksi yang benar dan salah pada masing-masing kelas. Untuk klasifikasi 2 kelas, komponen yang digunakan adalah

	Aktual Positif	Aktual Negatif
Prediksi Positif	TP	FP
Prediksi Negatif	FN	TN

Keterangan :

TP (True Positive) = data positif yang diprediksi positif.

TN (True Negative) = data negatif yang diprediksi negatif.

FP (False Positive) = data negatif yang diprediksi positif.

FN (False Negative) = data positif yang diprediksi negatif.

3.8 Perhitungan Accuracy

Accuracy digunakan untuk mengetahui tingkat ketepatan model dalam mengklasifikasikan data secara keseluruhan. Rumus accuracy yang digunakan adalah:

$$\text{Accuracy}= \frac{TP+TN}{TP+TN+FP+FN} \text{X100\%}$$

Pada klasifikasi lebih dari dua kelas, prinsipnya adalah menghitung jumlah prediksi yang benar dibandingkan dengan seluruh data pengujian. Jumlah data akhir yang digunakan adalah 1.617 data. Pembagian data dilakukan menjadi 90% data latih dan 10% data uji, 80% data latih dan 20% data uji, serta 70% data latih dan 30% data uji.

3.8.1 Skenario 90% data latih dan 10% data uji

Jumlah data uji:

$$1.617 \times 10\% = 161{,}7 = 162$$

Sehingga:

$$\text{Data latih} = 1.617 - 162 = 1.455$$

Pada pengujian 3 kelas, naskah melaporkan accuracy sebesar 98,76%. Jika dikonversikan ke jumlah prediksi benar:

$$98{,}76\% \times 162 = 159{,}991 = 160$$

Maka terdapat sekitar 160 data yang diklasifikasikan dengan benar dari 162 data pengujian. Perhitungannya:

$$\text{Accuracy} = \frac{160}{162} \text{ x } 100\%$$

$$\text{Accuracy} = 98{,}7654\%$$

$$\text{Accuracy} = 98{,}77\%$$

Jadi, accuracy pada skenario 90:10 adalah 98,77%.

3.8.2 Skenario 80% data latih dan 20% data uji

Jumlah data uji:

$$1.617 \times 20\% = 323{,}4 = 324$$

Jumlah data latih:

$$1.617 - 324 = 1.293$$

Jumlah prediksi benar :

$$324 \times 97{,}83\% = 316{,}9692 = 317$$

Sehingga perhitungan accuracy:

$$\text{Accuracy} = \frac{317}{324} \text{ x } 100\%$$

$$\text{Accuracy} = 97{,}8395\%$$

$$\text{Accuracy} = 97.84\ \%$$

Jadi, dari 324 data pengujian terdapat sekitar 317 data yang diklasifikasikan dengan benar.

3.8.3 Skenario 70% data latih dan 30% data uji

Jumlah data uji:

$$1.617 \times 30\% = 485{,}1 = 485$$

Jumlah prediksi benar =480

$$\text{Accuracy} = \frac{480}{485} \text{ x } 100\%$$

$$\text{Accuracy} = 98{,}97\%$$

3.9 Perhitungan Precision

3.9.1 Perhitungan Precision 2 Kelas

$$\text{Precision} = \frac{TP}{TP+FP} \text{X}100\%$$

TP = 477

FP = 4

$$\text{Precision} = \frac{477}{477+4} \text{X}100\%$$

$$\text{Precision} = \frac{477}{481} \text{X}100\% = 99{,}17\%$$

3.9.2 Perhitungan precision 3 kelas

$$\text{Precision} = \frac{TP}{TP+FP} \text{X}100\%$$

TP = 319

FP = 4

$$\text{Precision} = \frac{319}{319+4} \text{X}100\%$$

$$\text{Precision} = \frac{319}{323} \text{X}100\% = 98{,}76\%$$

3.10 Perhitungan Recall

3.10.1 Perhitungan recall 2 Kelas

Rumus:

$$\text{Recall} = \frac{TP}{TP+FP} \text{X}100\%$$

TP = 477

FP = 4

$$\text{Recall} = \frac{477}{477+4} \text{X}100\%$$

$$\text{Recall} = \frac{477}{481} \text{X}100\% = 99{,}17\%$$

3.10.2 Perhitungan recall 3 Kelas

Rumus:

$$\text{Recall} = \frac{TP}{TP+FP} \text{X}100\%$$

TP = 477

FP = 4

$$\text{Recall} = \frac{477}{477+4} \text{X100\%}$$

$$\text{Recall} = \frac{477}{481} \text{X100\%} = 99{,}17\%$$

3.11 Perhitungan F1-Score

3.11.1 Perhitungan 2 Kelas

Rumus:

$$F1=2\times\frac{PrecisionxRecall}{Precision+Recall}$$

Masukkan nilai:

$$F1=2\times\frac{0{,}9917x0{,}9917}{0{,}9917+0{,}9917}$$

Jadi secara matematis:

$$F1=0{,}9917 = 99{,}17\%$$

3.11.2 Perhitungan 3 Kelas

Rumus:

$$F1=2\times\frac{PrecisionxRecall}{Precision+Recall}$$

Masukkan nilai:

$$F1=2\times\frac{0{,}9876x0{,}9917}{0{,}9876+0{,}9917}$$

Jadi secara matematis:

$$F1=0{,}9896 = 98{,}96\%$$

Tabel hasil akhir

Klasifikasi	Accuracy	Precision	Recall	F1-Score
2 Kelas	99,17%	99,17%	99,17%	99,17%
3 Kelas	98,76%	98,76%	99,17%	98,96%

4 HASIL DAN PEMBAHASAN

Berdasarkan hasil pengolahan data, dari 2.363 data awal diperoleh 1.617 data yang dapat digunakan setelah proses preprocessing dan pembersihan data. Data tersebut terdiri dari 1.595 sentimen negatif, 15 sentimen positif, dan 7 sentimen netral. Selanjutnya, data diubah menjadi bentuk numerik menggunakan pembobotan TF-IDF dan diklasifikasikan menggunakan algoritma C4.5.

Pengujian dilakukan menggunakan tiga skenario pembagian data, yaitu 90%:10%, 80%:20%, dan 70%:30%. Hasil pembagian masing-masing menghasilkan data latih dan data uji sebanyak 1.455:162, 1.293:324, dan 1.132:485.

Hasil evaluasi menunjukkan bahwa pada klasifikasi dua kelas diperoleh accuracy sebesar 99,17%, precision sebesar 99,17%, recall sebesar 99,17%, dan F1-score sebesar 99,17%. Sementara itu, pada klasifikasi tiga kelas diperoleh accuracy sebesar 98,76%, precision sebesar 98,76%, recall sebesar 99,17%, dan F1-score sebesar 98,96%.

Hasil tersebut menunjukkan bahwa algoritma C4.5 mampu melakukan klasifikasi sentimen dengan nilai evaluasi yang tinggi pada data penelitian. Namun, jumlah data yang tidak seimbang, terutama dominasi sentimen negatif, menjadi salah satu hal yang perlu diperhatikan dalam menginterpretasikan hasil klasifikasi. Dengan demikian, penggunaan precision, recall, dan F1-score diperlukan sebagai pelengkap accuracy untuk memberikan gambaran hasil klasifikasi.

5 KESIMPULAN

Berdasarkan hasil penelitian, algoritma C4.5 dengan pembobotan TF-IDF dapat digunakan untuk melakukan analisis sentimen terhadap data media sosial mengenai penggundulan hutan dan bencana alam. Pada klasifikasi dua kelas diperoleh accuracy sebesar 99,17%, sedangkan pada klasifikasi tiga kelas diperoleh accuracy sebesar 98,76% dengan F1-score sebesar 98,96%. Hasil tersebut menunjukkan bahwa metode yang digunakan mampu melakukan klasifikasi terhadap data penelitian dengan hasil evaluasi yang tinggi. Namun, ketidakseimbangan jumlah data antar kelas masih menjadi keterbatasan yang perlu diperhatikan pada penelitian selanjutnya.

REFERENSI

[1] Murthy, D. (2024). Sociology of Twitter/X: Trends, Challenges, and Future Research Directions. Annual Review of Sociology, 50, 169–190.

[2] Ramadhan, C., Ruslanjari, D., & Nurjani, E. (2023). Spatial and temporal based deforestation proclivity analysis on flood events with applying watershed scale (case study: Lasolo watershed in Southeast Sulawesi, Central Sulawesi, and South Sulawesi, Indonesia). International Journal of Disaster Risk Reduction, 93, 103745..

[3] Ohtani, S. (2024). Deepening and Practical Application of Sentiment Analysis: Through Exploration of Public Interest and Sentiment on "Biodiversity" From Twitter. Environment and Behavior, 56(7–8).

[4] Hidayat, R., Hakim, A. B., & Nugraha, R. (2024). Perbandingan Metode Naïve Bayes dan Decision Tree C4.5 untuk Analisis Sentimen Produk Es Teh Indonesia di Media Sosial Twitter. Jurnal SISKOM-KB, 7(2)..

[5] Subandi, S., Nugroho, A. S. B., & Hidayat, M. (2026). Web-based sentiment analysis of environmental issues on social media X: A comparison of SVM and Random Forest. Journal of Soft Computing Exploration, 7(2).

[6] Jim, J. R., Talukder, M. A. R., Malakar, P., Kabir, M. M., Nur, K., & Mridha, M. F. (2024). Recent advancements and challenges of NLP-based sentiment analysis: A state-of-the-art review. Natural Language Processing Journal, 6, 100059.