E-ISSN : 3025-1311
https://ejournal.warunayama.org/kohesi

Kohesi: Jurnal Multidisiplin Saintek
Volume 7 No 6 Tahun 2025

PERBANDINGAN DECISION TREE DAN NAÏVE BAYES CLASSIFIER PADA ANALISIS SENTIMEN MENGENAI KEBIJAKAN PENGHAPUSAN KEWAJIBAN SKRIPSI

Ilham Ainur Idhana[1*], Basuki Rahmat[2], Eva Yulia Puspaningrum[3]
Universitas Pembangunan Nasional Veteran Jawa Timur, Indonesia
Jl. Rungkut Madya, Gn. Anyar, Kec. Gn. Anyar, Surabaya, Jawa Timur 60294
Email: 19081010099@student.upnjatim.ac.id[1*], basukirahmat.if@upnjatim.ac.id[2],
evapuspaningrum.if@upnjatim.ac.id[3]

Abstrak

Kebijakan penghapusan kewajiban skripsi sebagai syarat kelulusan mahasiswa S1/D4 yang diatur dalam Permendikbudristek Nomor 53 Tahun 2023 telah memunculkan berbagai reaksi di kalangan masyarakat, terutama di media sosial seperti Twitter. Oleh karena itu, algoritma Decision Tree (DT) Naïve Bayes Classifier (NBC) dan diterapkan untuk mengklasifikasikan sentimen masyarakat terhadap kebijakan penghapusan kewajiban skripsi di Indonesia. Tujuan utama dari penelitian ini adalah untuk mengevaluasi kinerja algoritma DT dan NBC dalam mengklasifikasikan data teks ke dalam tiga kategori, yaitu positif, netral, dan negatif. Dataset yang digunakan berupa ribuan tweet yang telah melalui tahap preprocessing dan pelabelan menggunakan pendekatan lexicon-based serta ekstraksi fitur menggunakan *TF-IDF*. Hasil penelitian menunjukkan bahwa sentimen negatif lebih dominan dibandingkan sentimen positif dan netral. Evaluasi performa model dilakukan menggunakan metrik akurasi, di mana hasil terbaik diperoleh pada model Naïve Bayes dengan akurasi 77.13% dan nilai rata-rata AUC-ROC sebesar 56%, sementara Decision Tree mencapai akurasi 74.21% dan nilai rata-rata AUC-ROC 53%. Hasil ini menunjukkan bahwa Naïve Bayes lebih unggul dalam klasifikasi teks berbasis opini masyarakat di media sosial dibandingkan dengan Decision Tree.

Kata Kunci: Analisis Sentimen, *Decision Tree, Naïve Bayes Classifier*, Twitter

Abstract

The policy of removing the thesis obligation as a graduation requirement for S1 / D4 students stipulated in Permendikbudristek Number 53 of 2023 has raised various reactions among the public, especially on social media such as Twitter. Therefore, the Decision Tree (DT) Naïve Bayes Classifier (NBC) algorithm is applied to classify public sentiment towards the policy of removing the thesis obligation in Indonesia. The main objective of this research is to evaluate the performance of DT and NBC algorithms in classifying text data into three categories, namely positive, neutral, and negative. The dataset used is thousands of tweets that have gone through

Article History
Received: March 2025
Reviewed: March 2025
Published: March 2025

Plagirism Checker No 234
Prefix DOI : Prefix DOI :
10.8734/Kohesi.v1i2.365
Copyright : Author
Publish by : Kohesi

preprocessing and labeling stages using a lexicon-based approach and feature extraction using TF-IDF. The results show that negative sentiment is more dominant than positive and neutral sentiment. Model performance evaluation was conducted using accuracy metrics, where the best results were obtained in the Naïve Bayes model with an accuracy of 77.13% and an average AUC-ROC value of 56%, while Decision Tree achieved an accuracy of 74.21% and an average AUC-ROC value of 53%. These results show that Naïve Bayes is superior in the classification of public opinion-based text on social media compared to Decision Tree.
***Keywords**: Sentiment Analysis, Decision Tree, Naïve Bayes Classifier, Twitter*

PENDAHULUAN

Skripsi sebagai tugas akhir mahasiswa merupakan topik yang banyak dibahas dalam literatur akademik. Kebijakan penghapusan kewajiban skripsi sebagai syarat kelulusan mahasiswa S1/D4 yang diatur dalam Permendikbudristek Nomor 53 Tahun 2023 tentang Penjaminan Mutu Pendidikan Tinggi telah menimbulkan perdebatan di kalangan akademisi, mahasiswa, dan masyarakat umum. Kebijakan ini memberikan fleksibilitas bagi perguruan tinggi untuk menentukan bentuk tugas akhir mahasiswa, seperti proyek, prototipe, atau bentuk lainnya yang lebih relevan dengan bidang studi masing-masing. Beberapa pihak mendukung kebijakan ini karena dianggap dapat meningkatkan inovasi dan mengurangi beban administratif bagi mahasiswa. Namun, tidak sedikit pula yang mengkritiknya karena khawatir akan menurunkan standar kualitas lulusan. Reaksi pro dan kontra ini banyak muncul di media sosial, terutama di Twitter, yang berfungsi sebagai platform utama bagi masyarakat untuk mengekspresikan opini mereka secara bebas dan real-time. Oleh karena itu, diperlukan analisis sentimen untuk memahami pola kecenderungan opini publik terhadap kebijakan ini.

Analisis sentimen telah banyak digunakan dalam berbagai penelitian untuk mengkaji opini masyarakat terhadap suatu kebijakan, produk, atau peristiwa tertentu. Berbagai metode telah digunakan dalam analisis sentimen, termasuk pendekatan berbasis *machine learning* seperti *Naïve Bayes Classifier (NBC)* dan *Decision Tree (DT)*. *Naïve Bayes* adalah metode berbasis probabilitas yang banyak digunakan dalam klasifikasi teks dan telah terbukti efektif dalam melakukan analisis sentimen.. Penelitian yang dilakukan oleh (Kaburuan et al., 2022) terhadap ulasan produk di marketplace Shopee menunjukkan bahwa model *Naïve Bay*es memberikan akurasi tinggi dalam mengklasifikasikan sentimen pelanggan. Penelitian lain juga oleh (Saputra et al., 2023) mengenai Piala Dunia Fifa 2022 juga didapatkan bahwa model *NBC* dapat mengklasifikasi data *tweet* dari Twitter dengan akurasi tinggi. Di sisi lain, *Decision Tree* bekerja dengan membangun model berbasis aturan yang dapat mengklasifikasikan data dengan struktur pohon keputusan. Studi yang dilakukan oleh (Rizkia et al., 2019.) menunjukkan bahwa *Decision T*ree juga memberikan hasil akurasi tinggi dalam analisis sentimen terhadap penyedia layanan internet. Penelitian lain juga dilakukan oleh (Huda et al., 2023) terhadap layanan jasa pengiriman kurir pada ulasan *Play Store* mendapatkan hasil klasifikasi yang tinggi.

Perbandingan antara *Naïve B*ayes dan *Decision T*ree dalam analisis sentimen telah banyak dikaji dalam berbagai konteks. Beberapa penelitian menunjukkan bahwa *Naïve Baye*s lebih

unggul dalam menangani teks karena sifatnya yang probabilistik dan toleran terhadap data yang tidak seimbang. Namun, *Decision T*ree sering kali memberikan interpretasi yang lebih jelas dan mudah dipahami dalam proses pengambilan keputusan. Meskipun demikian, belum ada kajian yang secara spesifik membandingkan kedua algoritma ini dalam analisis sentimen mengenai kebijakan penghapusan kewajiban skripsi sebagai syarat lulus mahasiswa SI/D4 di Indonesia.

Berdasarkan latar belakang tersebut, penelitian ini bertujuan untuk mengevaluasi dan membandingkan performa Naïve Bayes serta Decision Tree dalam menganalisis sentimen terhadap kebijakan penghapusan kewajiban skripsi di Twitter. Hipotesis awal yang diajukan adalah bahwa *Naïve Bayes* akan memiliki akurasi yang lebih tinggi dibandingkan *Decision Tre*e dalam klasifikasi sentimen, mengingat karakteristiknya yang lebih sesuai untuk analisis teks dan lebih tahan terhadap noise dalam data media sosial. Melalui penelitian ini, diharapkan dapat diperoleh wawasan yang lebih mendalam mengenai kecenderungan opini masyarakat serta efektivitas setiap algoritma dalam mengklasifikasikan sentimen berbasis teks.

KAJIAN TEORITIS

Berbagai penelitian telah dilakukan dengan menerapkan algoritma machine learning dalam analisis sentimen untuk mengkaji opini masyarakat terhadap suatu kebijakan atau isu tertentu. Kaburuan et al. (2022) menggunakan Naïve Bayes Classifier pada analisis sentimen ulasan produk pakaian wanita di Shopee. Saputra et al. (2023) menggunakan Naïve Bayes Classifier dan Support Vector Machine sebagai perbandingan pada analisis sentimen masyarakat twitter terhadap Piala Dunia FIFA 2022. Rizkia et al. (2019) menggunakan Decision Tree pada analisis sentiment pelanggan terhadap layanan Indihome di Twitter. Hanafi & Solichin (2023) menggunakan Multinomial Naïve Bayes pada analisis sentimen masyarakat terhadap PSSI terkait tragedi kanjuruhan.

Analisis sentimen, atau yang dikenal juga sebagai opinion mining, merupakan studi komputasi yang bertujuan untuk mengenali dan mengekspresikan opini, sentimen, evaluasi, sikap, emosi, subjektivitas, penilaian, atau pandangan yang terkandung dalam sebuah teks (Jindal & Liu, 2008). Analisis sentimen mengelompokkan kalimat atau opini berdasarkan kecenderungan popularitas teks yang terkandung di dalamnya. Popularitas ini mencerminkan pendapat dengan aspek positif yang berkaitan dengan hal baik atau kondisi normal, serta aspek negatif yang berhubungan dengan hal buruk atau situasi yang tidak diinginkan. Dalam konteks Twitter, analisis sentimen berfungsi sebagai alat untuk mengevaluasi persepsi masyarakat dengan fokus pada identifikasi sentimen dalam setiap tweet individu.

Decision Tree

Decision Tree adalah algoritma yang bekerja dengan membangun model berbasis aturan yang menyerupai struktur pohon. Setiap node dalam pohon mewakili fitur tertentu, sedangkan cabang mewakili keputusan berdasarkan fitur tersebut. Algoritma Decision Tree bekerja dengan mengubah data menjadi serangkaian aturan keputusan (Salman, 2020). Salah satu algoritma yang secara khusus digunakan dalam pembentukan Decision Tree adalah ID3 (Iterative Dichotomiser 3). Algoritma ID3 didasarkan pada konsep Entropy dan Information Gain. Nilai Entropy dapat dihitung menggunakan rumus pada persamaan 2.1..

$$Entropy\ (S) = \sum_{i=1}^{k} - p_i * log_2\ p_i \qquad (2.1)$$

Keterangan:

S = ruang (data) *sample* yang digunakan untuk *training*

k = jumlah partisi pada S
p_i = probabilitas dari kelas ke-i (jumlah sampel kelas ke-i dibagi total sampel)
Untuk nilai *Information Gain* (*IG*) dapat ditemukan dengan menggunakan rumus berikut

$$Gain\ (A) = Entropy\ (S) - \sum_{i=1}^{k} \frac{|S_1|}{|S|} xEntropy\ (S_i) \quad (2.2)$$

Keterangan:
S : Ruang (data) *sample* yang digunakan untuk *training*
A : Atribut
$|S_1|$: Jumlah *sample* untuk nilai V
$|S|$: Jumlah seluruh *sample data*
$Entropy\ (S)$: *entropy* untuk *sample-sample* yang memiliki nilai i

Naïve Bayes

Klasifikasi Bayesian merupakan metode klasifikasi data berbasis model statistik yang memungkinkan perhitungan probabilitas keanggotaan suatu kelas. Metode NBC berdasar pada teorema Bayes untuk peluang bersyarat. Teorema ini menghitung peluang suatu pengamatan masuk dalam kategori (C_j) dengan syarat kondisi variabel lainnya (X) terpenuhi

$$P(C_j|X) = \frac{P(X|C_j\}.P(C_j\}}{P(X)} \quad (2.3)$$

Keterangan:
$P(C_j|X)$: Probabilitas pengamatan berkategori C_j berdasarkan kondisi X (*posterior probability).*
$P(X|C_j)$: Probabilitas X dengan kemungkinan C_j (*likelihood*).
$P(C_j)$: Nilai probabilitas pada data *training* melalui perhitungan kategori C_j (*prior*).
$P(X)$: Jumlah probabilitas *features* data *training* X ($x_1, x_2, \dots, x_k$) (evidence)

METODE PENELITIAN

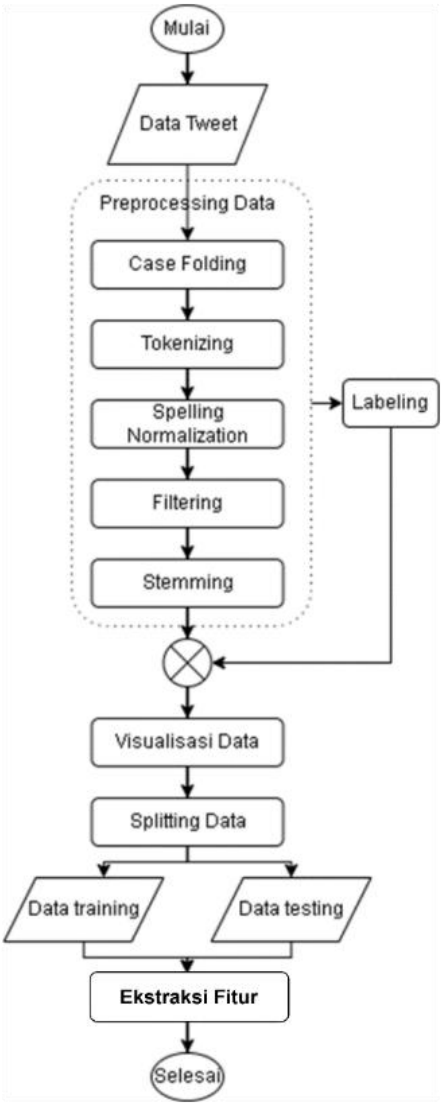
Bab ini akan menjelaskan tahapan sistematis atau terstruktur yang digunakaan dalam penelitian akan dilaksanakan, serta akan menjelaskan runtutan proses penelitian dari pengumpulan data, pengolahan data, pembuatan model klasifikasi, pengujian *tuning hyperparameter* dan interpretasi hasil ujian.

Pengumpulan Data

Data dikumpulkan melalui proses crawling dan scraping dari platform media sosial Twitter menggunakan Tweet Harvest dalam pemrograman Python. Dengan memanfaatkan TwitterAPI, tweet yang berkaitan dengan kebijakan penghapusan kewajiban skripsi sebagai syarat kelulusan di Indonesia diperoleh berdasarkan kata kunci yang meliputi "skripsi dihapus", "kebijakan skripsi", "skripsi diganti", "tanpa skripsi", dan "permendikbud 53". Tweet yang dikumpulkan berasal dari periode 30 Agustus 2023 hingga 30 Desember 2024. Pemilihan tanggal awal tersebut didasarkan pada waktu awal munculnya isu terkait kebijakan penghapusan kewajiban skripsi.

Pengolahan Data

Preprocessing data diperlukan untuk mengubah data mentah menjadi format yang lebih terorganisir dan bermakna, sehingga mendukung proses klasifikasi dan mengoptimalkan kinerja model, terutama dalam meningkatkan akurasi. (Damaratih, 2021).



Gambar 1. Diagram Alir Pengolahan Data

Tahap dari *preprocessing* meliputi *case folding, tokenizing, spelling normalization, filtering,* dan *stemming.* Tahap *labeling* pada ribuan data tweet menggunakan *lexicon-based Indonesia.* Tahap ekstraksi fitur menggunakan teknik *TF-IDF* yang menghasilkan *Document-Term Matrix (DTM).*

Pembuatan Model Klasifikasi Decision Tree

Membentuk model klasifikasi atau regresi dalam bentuk struktur pohon yang pada prosesnya melibatkan pembagian data menjadi bagian-bagian yang lebih kecil dan secara bertahap membangun struktur pohon keputusan. Pohon ini terdiri dari simpul keputusan dan simpul daun. Secara umum proses klasifikasi algoritma *Decision Tree* adalah sebagai berikut.

1. Pemilihan akar pohon didasarkan pada atribut dengan nilai gain tertinggi. Setelah menghitung nilai gain dari setiap atribut, atribut dengan nilai tertinggi akan dijadikan sebagai akar pohon utama.
2. Sebelum menentukan nilai gain pada suatu atribut, langkah pertama yang dilakukan adalah menghitung nilai entropy menggunakan Persamaan 2.1.
3. Hitung nilai *gain* menggunakan rumus pada Persamaan 2.2
4. Mengulangi Langkah ke-2 hingga semua data (*record*) terbagi ke dalam partisi yang sesuai.
5. Proses pemisahan (partisi) pohon akan berhenti ketika:
   a. Seluruh data (record) dalam simpul N memiliki kelas yang sama.

b. Tidak ada lagi atribut dalam data (record) yang dapat dipartisi lebih lanjut.
c. Tidak ada cabang dengan data (record) yang kosong.

Pembuatan Model Klasifikasi Naïve Bayes

Pendekatan Naïve Bayes yang diterapkan dalam penelitian ini adalah Multinomial Naïve Bayes, karena lebih sesuai untuk menangani fitur berupa bilangan diskrit, seperti frekuensi kemunculan kata yang diperoleh dari hasil document-term matrix (Singh et al., 2019). Rumus umum Naïve Bayes, seperti yang ditunjukkan pada Persamaan 1, mendefinisikan P(C_j | X) sebagai probabilitas posterior, yaitu probabilitas suatu pengamatan termasuk dalam kategori C_j dengan kondisi X. Komponen P(X|C_j) merepresentasikan *likelihood*, yaitu probabilitas X dengan asumsi kategori C_j. Nilai ini kemudian dikalikan dengan P(C_j), yang merupakan probabilitas prior atau probabilitas kategori C_j dalam data *training*. Hasil perkalian tersebut dibagi dengan P(X), yaitu *evidence* atau total probabilitas semua fitur data *training* X (x1, x2, …, xk). Dengan demikian, pendekatan Multinomial dapat dinotasikan sebagai berikut.

$$P(x_1 = a_1, \dots, x_1 = a_k | C = c_j) \approx \frac{L!}{\prod_{i=1}^{k} a_i!} \prod_{i=1}^{k} P(x_i | C = c_j) \quad (2.4)$$

Dengan variabel L merepresentasikan banyak kata dalam satu dokumen atau *tweet* dan a_i merepresentasikan jumlah kata ke-i (x_i) dalam satu *tweet.*

Pengujian Tuning Hyperparameter

Pada tahap ini, dilakukan pengujian untuk menentukan kondisi optimal dari model klasifikasi Naïve Bayes dan Decision Tree berdasarkan hasil tuning hyperparameter. Pengujian dilakukan dengan mengombinasikan berbagai rasio pembagian data dan nilai-nilai hyperparameter. Rasio pembagian data yang diuji mencakup 70:30, 80:20, dan 90:10, dengan porsi terbesar dialokasikan untuk data training dan sisanya untuk data testing. Pada Decision Tree, nilai minimum sample split yang digunakan meliputi 2, 10, 15, dan 20, sementara nilai minimum sample leaf mencakup 1, 5, 10, dan 20. Untuk Naïve Bayes, nilai alpha atau laplace smoothing yang diuji mencakup 1, 2, 5, dan 10.. Selain itu juga digunakan metode K-Fold Cross Validation sebanyak 5-fold sebagai cara untuk validasi kemampuan model dan mengidentifikasi konsistensi model pada DT dan NBC dari hasil tuning hyperparameter.

Interpretasi Hasil Pengujian

Evaluasi performa dilakukan berdasarkan nilai accuracy, recall, sensitivity, dan specificity yang diperoleh dari confusion matrix. Penelitian ini menentukan model yang lebih unggul dengan menganalisis kurva Receiver Operating Characteristic (ROC) dan nilai Area Under Curve (AUC). Analisis AUC-ROC bertujuan untuk mengukur sejauh mana model mampu membedakan kelas berdasarkan berbagai nilai threshold. Untuk mengevaluasi model dengan multi-kelas, digunakan metode one-vs-rest (OvR), di mana setiap kelas secara bergantian dianggap sebagai kelas positif sementara yang lain sebagai negatif. Pada tahap akhir penelitian, dilakukan interpretasi terhadap hasil performa kedua model.

HASIL DAN PEMBAHASAN

Bagian ini menyajikan hasil penelitian yang bertujuan untuk menganalisis sentimen masyarakat terhadap kebijakan penghapusan kewajiban skripsi sebagai syarat kelulusan mahasiswa S1/D4 di Indonesia. Hasil yang diperoleh mencakup tahapan utama, mulai dari

pengumpulan data, pengolahan data, pengujian tuning hyperparameter, hingga interpretasi hasil pengujian.

Pengumpulan Data

Penelitian ini menggunakan dataset berupa tweet dari Twitter yang membahas tentang "kebijakan penghapusan kewajiban skripsi." Data dikumpulkan dengan teknik crawling menggunakan TweetHarvest, menghasilkan sebanyak 2482 data. Namun, dataset tersebut masih dalam bentuk mentah, yang mengandung banyak informasi tidak relevan serta atribut yang kurang mendukung analisis.

	conversation_id_str	created_at	favorite_count	full_text	id_str	image_url	in_reply_to_screen_name	lang	location	quote_count	reply_count	retweet_count	tweet_url	user_id_str	username
0	1793945602422421302	Fri May 24 09:59:58 +0000 2024	0	cek tumblr / parafrase / makalah / skripsi -	1793945602422421302	https://pbs.twimg.com/media/GOVfeTpaEAASSRx.jpg	NaN	in	she/her	0	0	0	https://x.com/cizarine/status/1793945602422421302	1199033824793796608	cizarine
1	1793888687094452654	Fri May 24 06:16:06 +0000 2024	1	cek tumblr / parafrase / makalah / skripsi -	1793888687094452654	https://pbs.twimg.com/media/GOUsPpoaAAQZrxU.jpg	NaN	in	she/her	0	0	0	https://x.com/cizarine/status/1793888687094452654	1199033824793796608	cizarine
2	1793873302508982350	Fri May 24 05:14:58 +0000 2024	0	Wait... statistika inferensial ilu yg buat pe...	1793873302508982350	NaN	NaN	in	Indonesia	0	0	1	https://x.com/chordream/status/1793873302508982...	1306637875136442369	chordream
3	1793814476912996542	Fri May 24 01:21:13 +0000 2024	1	cek tumblr / parafrase / makalah / skripsi -	1793814476912996542	https://pbs.twimg.com/media/GOTowLMbwAEfsLcS.jpg	NaN	in	she/her	0	0	0	https://x.com/cizarine/status/1793814476912996542	1199033824793796608	cizarine
4	1793595154672527126	Thu May 23 10:49:42 +0000 2024	0	cek tumblr / parafrase / makalah / skripsi -	1793595154672527126	https://pbs.twimg.com/media/GOQhR1PaYAAig1o7.jpg	NaN	in	she/her	0	0	0	https://x.com/cizarine/status/1793595154672527126	1199033824793796608	cizarine
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
759	1696915134227292264	Thu Aug 31 01:35:58 +0000 2023	0	@collegemenfess bukan dihapus ga si tp digant...	1697060628673019930	NaN	collegemenfess	in	NaN	0	0	0	https://x.com/ssh1ttluv/status/1697060628673...	1424225795854114820	ssh1ttluv
760	1697059649361793348	Thu Aug 31 01:32:54 +0000 2023	0	mending skripsi yang dihapus anjir. kalo diedit...	1697059649361793348	NaN	NaN	in	💗💚	0	0	0	https://x.com/gogobugi/status/1697059649361679...	1319412450119147521	gogobugi
761	1697050046944694279	Thu Aug 31 01:25:42 +0000 2023	1	Klo dosbing dihapus etu mau ngerjain skripsi d...	1697050046944694279	NaN	NaN	in	NaN	0	0	0	https://x.com/jaithype/status/1697050046944694279	1476371700421857285	jaithype
762	1697057864967431020	Thu Aug 31 01:24:59 +0000 2023	0	Beneran kagak ni? Otw nyusun ni btw kalo diter...	1697057864967431020	NaN	NaN	in	NaN	0	0	0	https://x.com/naqbail/status/1697057864967431020	1131124706583633920	naqbail
763	1697054557589967065	Thu Aug 31 01:28:15 +0000 2023	0	@mukimagoreng Nggak dihapus gasih itu om clickb...	1697059672546775549	NaN	mukimagoreng	in	She 🏳 AR 60	0	1	0	https://x.com/ktoorush/status/1697059672546775549	1352263140138883314	ktoorush

764 rows × 15 columns

Gambar 2. Hasil Crawling Data

Pengolahan Data

Data *tweet* yang diambil dari platform Twitter merupakan data yang tidak terstruktur, sehingga diperlukan proses dalam text mining, yaitu preprocessing data, untuk meningkatkan akurasi klasifikasi. Langkah-langkah yang dilakukan dalam preprocessing meliputi case folding, tokenisasi, normalisasi ejaan, filtering, dan stemming. Tahap selanjutnya melibatkan proses pelabelan data dan ekstraksi fitur. Ekstraksi fitur dilakukan menggunakan metode *document-term matrix*.

	Sentimen	tweet_cleaned
0	@radityaprkz Tahun depan skripsi dihapus pak!!	tahun depan skripsi dihapus pak
1	@lucelyf_ Kabarnya kemaren skripsi dihapus tauuu	kabarnya kemaren skripsi dihapus tauuu
2	@collegemenfess Skripsi dihapus	skripsi dihapus
3	@collegemenfess Skripsi salah satu hal yg meng...	skripsi salah satu hal yg mengajarkan mahasis...
4	Bisaga skripsi dihapus anyinggg ak tidak sangg...	bisaga skripsi dihapus anyinggg ak tidak sangg...

Gambar 3. Hasil Casefolding

	Sentimen	tweet_cleaned	tokenized
0	@radityaprkz Tahun depan skripsi dihapus pak!!	tahun depan skripsi dihapus pak	[tahun, depan, skripsi, dihapus, pak]
1	@lucelyf_ Kabarnya kemaren skripsi dihapus tauuu	kabarnya kemaren skripsi dihapus tauuu	[kabarnya, kemaren, skripsi, dihapus, tauuu]
2	@collegemenfess Skripsi dihapus	skripsi dihapus	[skripsi, dihapus]
3	@collegemenfess Skripsi salah satu hal yg meng...	skripsi salah satu hal yg mengajarkan mahasis...	[skripsi, salah, satu, hal, yg, mengajarkan, m...
4	Bisaga skripsi dihapus anyinggg ak tidak sangg...	bisaga skripsi dihapus anyinggg ak tidak sangg...	[bisaga, skripsi, dihapus, anyinggg, ak, tidak...

Gambar 4. Hasil Tokenizing

	Sentimen	tokenized	normalized
0	@radityaprkz Tahun depan skripsi dihapus pak!!	[tahun, depan, skripsi, dihapus, pak]	[tahun, depan, skripsi, dihapus, pak]
1	@lucelyf_ Kabarnya kemaren skripsi dihapus tauuu	[kabarnya, kemaren, skripsi, dihapus, tauuu]	[kabarnya, kemaren, skripsi, dihapus, tahu]
2	@collegemenfess Skripsi dihapus	[skripsi, dihapus]	[skripsi, dihapus]
3	@collegemenfess Skripsi salah satu hal yg meng...	[skripsi, salah, satu, hal, yg, mengajarkan, m...	[skripsi, salah, satu, hal, yang, mengajarkan,...
4	Bisaga skripsi dihapus anyinggg ak tidak sangg...	[bisaga, skripsi, dihapus, anyinggg, ak, tidak...	[bisa, skripsi, dihapus, anjing, aku, tidak, s...

Gambar 5. Hasil Spelling Normalization

	Sentimen	filtered	stemmed
0	@radityaprkz Tahun depan skripsi dihapus pak!!	[skripsi, dihapus]	[skripsi, hapus]
1	@lucelyf_ Kabarnya kemaren skripsi dihapus tauuu	[kabarnya, kemaren, skripsi, dihapus]	[kabar, kemaren, skripsi, hapus]
2	@collegemenfess Skripsi dihapus	[skripsi, dihapus]	[skripsi, hapus]
3	@collegemenfess Skripsi salah satu hal yg meng...	[skripsi, salah, mengajarkan, mahasiswa, analy...	[skripsi, salah, ajar, mahasiswa, analytical, ...
4	Bisaga skripsi dihapus anyinggg ak tidak sangg...	[skripsi, dihapus, anjing, sanggup, membayangkan]	[skripsi, hapus, anjing, sanggup, bayang]

Gambar 6. Hasil Stemming

	Sentimen	tweet_cleaned	preprocessed_tweet	token	sentimen_new
0	@radityaprkz Tahun depan skripsi dihapus pak!!	tahun depan skripsi dihapus pak	skripsi,hapus	['skripsi', 'hapus']	-1
1	@lucelyf_ Kabarnya kemaren skripsi dihapus tauuu	kabarnya kemaren skripsi dihapus tauuu	kabar,kemaren,skripsi,hapus,tauuu	['kabar', 'kemaren', 'skripsi', 'hapus', 'tauuu']	-1
2	@collegemenfess Skripsi dihapus	skripsi dihapus	skripsi,hapus	['skripsi', 'hapus']	-1
3	@collegemenfess Skripsi salah satu hal yg meng...	skripsi salah satu hal yg mengajarkan mahasis...	skripsi,salah,yg,ajar,mahasiswa,ttg,analytical...	['skripsi', 'salah', 'yg', 'ajar', 'mahasiswa'...	-1
4	Bisaga skripsi dihapus anyinggg ak tidak sangg...	bisaga skripsi dihapus anyinggg ak tidak sangg...	bisaga,skripsi,hapus,anyinggg,ak,sanggup,bayang	['bisaga', 'skripsi', 'hapus', 'anyinggg', 'ak...	1
...	...	...	...	...	...

Gambar 7. Hasil Labeling

```
datatrain = pd.DataFrame(XTrainVectorized, columns=features)
datatrain
```

[13] ✓ 0.0s

aaaargggh	aamiin	aarghhh	aataga	aawowkwk	abadi	abis	abisan	about	...	zamane	ziciw	zinc	zona	zonasi	zone	zoom
0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0

Gambar 8. Hasil Ekstraksi Fitur *document-term-matrix*

Hasil dari preprocessing data dan labeling data diakumulasikan jumlahnya dalam Tabel 1 seperti berikut.

Tabel 1. Akumulasi Jumlah Data Tiap Label

Sentimen Netral	Sentimen Positif	Sentimen Negatif
167	599	1460

Hasil Pengujian Tuning Hyperparameter Decision Tree

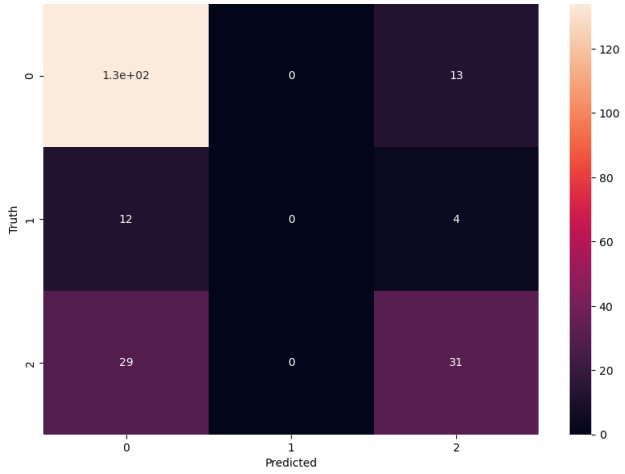
Hasil pengujian model *Decision Tree* dengan melibatkan variasi rasio pembagian data dan nilai *min split* dan nilai *min sample leaf* dituliskan melalui Tabel 2

Tabel 2. Hasil Pengujian *Tuning Hyperparameter Naïve Bayes*

		Min Split	Rasio Splitting Dataset		
			70:30	80:20	90:10
Min Leaf	1	2	67.81	67.26	68.60
		5	68.56	65.47	69.95
		10	67.96	67.93	69.55
		20	68.11	69.05	71.74
	5	2	68.71	68.16	71.30
		5	68.86	68.16	68.16
		10	67.81	69.28	70.85

		20	69.01	72.64	73.54
	10	2	69.61	72.64	73.09
		5	69.61	72.64	71.74
		10	69.61	72.64	73.54
		20	69.61	73.09	73.54
	20	2	71.85	74.21	73.99
		5	71.85	74.21	73.99
		10	71.85	74.21	73.99
		20	71.85	74.21	73.99

Berdasarkan uraian hasil akurasi pada Tabel 2. terlihat bahwa kombinasi nilai min samples split sebesar 20, min samples leaf sebesar 20 dan rasio splitting sebesar 80:20 memiliki tingkat akurasi tertinggi sebesar 74.21%. Untuk mengetahui tingkat overfitting, perlu mempertimbangkan nilai recall yang didapat melalui confusion matrix masing-masing model pengujian. Berikut tabel *confusion matrix* dari uji iterasi ke-38 yang ditampilkan pada Gambar 9.



	precision	recall	f1-score	support
-1	0.77	0.91	0.83	147
0	0.00	0.00	0.00	16
1	0.65	0.52	0.57	60
accuracy			0.74	223
macro avg	0.47	0.48	0.47	223
weighted avg	0.68	0.74	0.70	223

Gambar 9. Nilai *Confusion Matrix Model Terbaik DT*

Model tersebut menunjukkan keterbatasan dalam memprediksi sentimen netral (nilai *recall* 0) dan positif (nilai *recall* 0.52). Meskipun demikian, rata-rata nilai *recall* secara keseluruhan mencapai 0.74, yang menunjukkan performa yang cukup baik untuk model klasifikasi multi-kelas. Oleh karena itu, disimpulkan bahwa model Decision Tree (DT) dengan *min_samples_split* sebesar 20 dan *min_samples_leaf* sebesar 20 merupakan konfigurasi model terbaik untuk studi kasus analisis sentimen kebijakan penghapusan kewajiban skripsi di Indonesia.

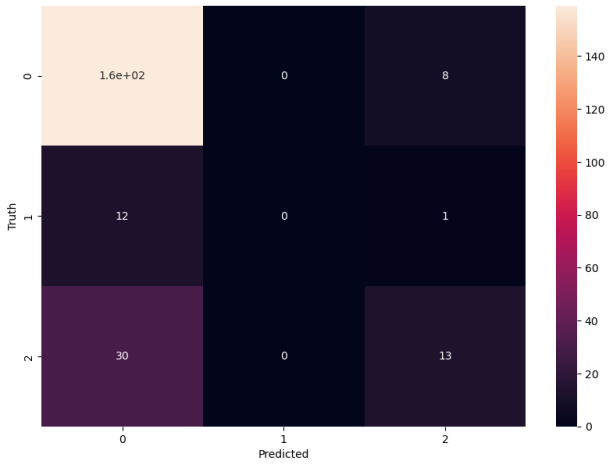
Hasil Pengujian Tuning Hyperparameter Naïve Bayes

Hasil pengujian model Multinomial Naïve Bayes, yang melibatkan variasi rasio pembagian data dan nilai *Laplace smoothing* atau *alpha*, disajikan dalam Tabel 3.

Tabel 3. Hasil Pengujian *Tuning Hyperparameter Naïve Bayes*

Alpha	Rasio *Splitting Dataset*		
	70:30	**80:20**	**90:10**
1	70.65%	70.85%	77.13%
2	69.61%	70.62%	76.68%
5	69.16%	71.30%	76.23%
10	69.01%	71.07%	75.33%

Berdasarkan uraian hasil akurasi pada Tabel 3 terlihat bahwa kombinasi nilai alpha yaitu 1 dan rasio splitting sebesar 90:10 memiliki tingkat akurasi tertinggi sebesar 77.13%. Untuk mengetahui tingkat overfitting, nilai rata-rata recall terhadap setiap prediksi kelas perlu dipertimbangkan. Nilai recall dapat diketahui melalui nilai confusion matrix masing-masing model pengujian. Berikut tabel *confusion matrix* dari uji iterasi ke-12 yang ditampilkan pada Gambar 10.



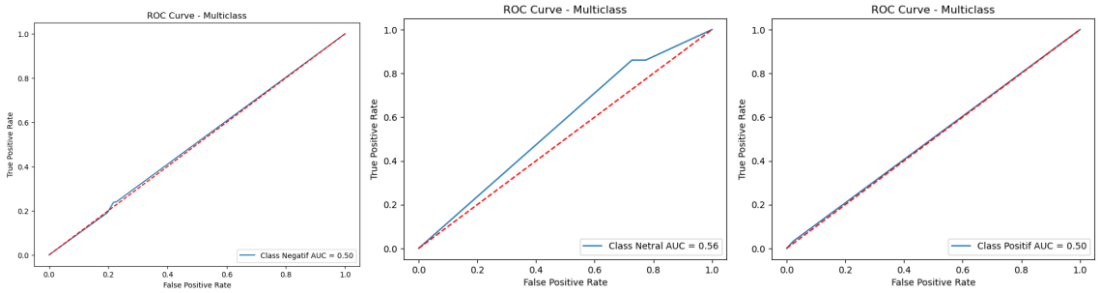
	precision	recall	f1-score	support
-1	0.79	0.95	0.86	167
0	0.00	0.00	0.00	13
1	0.59	0.30	0.40	43
accuracy			0.77	223
macro avg	0.46	0.42	0.42	223
weighted avg	0.71	0.77	0.72	223

Gambar 10. Nilai *Confusion Matrix* Model Terbaik *NBC*

Nilai *recall* untuk sentimen positif pada model tersebut tergolong rendah, yaitu hanya 0,3, yang mengindikasikan keterbatasan dalam memprediksi sentimen positif. Namun, secara keseluruhan, rata-rata nilai *recall* mencapai 0,77, yang menunjukkan performa yang cukup baik untuk model klasifikasi multi-kelas. Dengan demikian, dapat disimpulkan bahwa model Decision Tree (DT) dengan nilai *alpha* sebesar 1 dan rasio pembagian data 90:10 merupakan konfigurasi model terbaik untuk studi kasus analisis sentimen terhadap kebijakan penghapusan kewajiban skripsi di Indonesia.

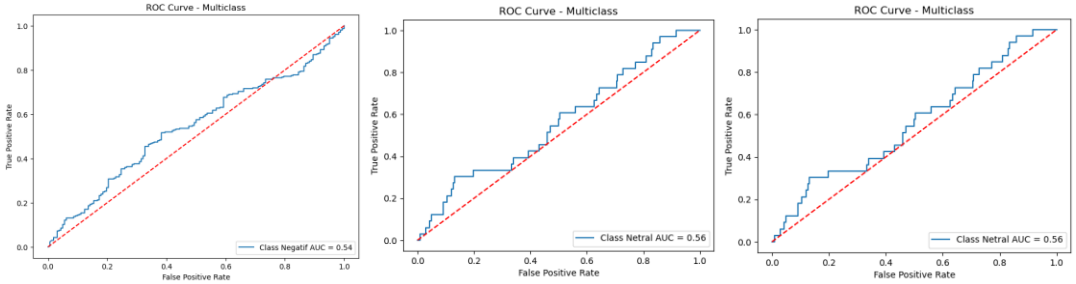
Interpretasi Performa Model Hasil Pengujian

Gambar 11 menampilkan kurva ROC dan nilai AUC yang dihasilkan oleh model Decision Tree dengan parameter sebagai berikut. rasio pembagian data *training* dan *testing* 80:20, nilai *min_samples_split=20*, serta *min_samples_leaf=2*.



Gambar 11. ROC AUC model Decision Tree

Gambar 12 menyajikan kurva ROC dan nilai AUC yang dihasilkan oleh model Multinominal Naïve Bayes dengan konfigurasi berikut: rasio pembagian data *training* dan *testing* sebesar 90:10, serta nilai *alpha* sebesar 1.



Gambar 12. ROC AUC model Naïve Bayes

Berdasarkan Gambar 11, model Decision Tree menghasilkan nilai AUC rata-rata sebesar 53%, sedangkan Naïve Bayes (NBC) pada Gambar 12 mencapai nilai AUC rata-rata 56%. Model NBC dengan nilai hyperparameter *alpha=1* dan rasio pembagian data 80:20 menunjukkan akurasi sebesar 77,13% serta skor AUC-ROC rata-rata 56%. Di sisi lain, model Decision Tree dengan rasio pembagian data 90:10 dan parameter *min_samples_split=20* serta *min_samples_leaf=2* menghasilkan nilai untuk akurasi 74,21% dan skor AUC-ROC rata-rata 53%. Naïve Bayes unggul dalam akurasi (+2,92%) dan AUC-ROC (+3%)

KESIMPULAN

Dapat disimpulkan bahwa Naïve Bayes memiliki kinerja yang lebih baik dibandingkan Decision Tree, dengan selisih akurasi mencapai 4,26% dan perbedaan rata-rata nilai AUC-ROC sebesar 3%. Naïve Bayes memperoleh akurasi sebesar 77,13% untuk pembagian data 90:10, dengan rata-rata skor AUC-ROC sebesar 56%. Nilai ini lebih tinggi dibandingkan dengan Decision Tree, yang pada pembagian data 80:20 menghasilkan akurasi sebesar 74,21% dan skor AUC-ROC sebesar 53%.

Hal ini membuktikan bahwa algoritma *Naive Bayes Classifier* lebih cocok terhadap data teks berbasis probabilitas dan diperkuat dengan dataset yang kecil sehingga lebih cepat dan efisien. Dengan demikian, dalam penelitian ini dapat disimpulkan bahwa metode klasifikasi *Naïve Bayes Classifier* memiliki kinerja lebih unggul dibandingkan *Decision Tree* dalam hal akurasi klasifikasi untuk tugas klasifikasi data teks pada studi kasus analisis sentiment masyarakat Indonesia terhadap kebijakan penghapusan kewajiban skripsi sebagai syarat kelulusan bagi mahasiswa S1/D4 di Indonesia.

DAFTAR REFERENSI

Aggarwal, C. C. (2015). *Data Mining*. Springer International Publishing. https://doi.org/10.1007/978-3-319-14142-8

Amirul Haj, A. S., Amrizal, V., & Arini, A. (2020). Analisis Sentimen Kinerja KPU Pemilu 2019 Menggunakan Algoritma K-Means Dengan Algoritma Confix Stripping Stemmer. *Journal of Innovation Information Technology and Application (JINITA)*, *2*(01), 9–18. https://doi.org/10.35970/jinita.v2i01.119

Huda, I., Nishfi, D., Prianto, C., & Awangga, R. M. (2023). *Dellavianti Nishfi Ilmiah Huda Analisis Sentimen Perbandingan Layanan Jasa Pengiriman Kurir Pada Ulasan Play Store Menggunakan Metode Random Forest dan Descision Tree*.

Hanafi, M. A., & Solichin, A. (2023). *ANALISIS SENTIMEN TERHADAP PSSI ATAS TRAGEDI KANJURUHAN MENGGUNAKAN MULTINOMIAL NAÏVE BAYES*. https://journal.budiluhur.ac.id/index.php/telematika/

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning***.** New York: springer.

Jindal, N., & Liu, B. (2008). *Opinion Spam and Analysis*. http://money.cnn.com/2006

Kaburuan, E. R., Sari, Y. S., & Agustina, I. (2022). Sentiment Analysis on Product Reviews from Shopee Marketplace using the Naïve Bayes Classifier. *Lontar Komputer : Jurnal Ilmiah Teknologi Informasi*, *13*(3), 150. https://doi.org/10.24843/lkjiti.2022.v13.i03.p02

Kelly, A., & Johnson, M. A., (2021) *"Investigating the Statistical Assumptions of Naïve Bayes Classifiers,"* 2021 55th Annual *Conference on Information Sciences and Systems (CISS)*, Baltimore, MD, USA, 2021, pp. 1-6, doi: 10.1109/CISS50987.2021.9400215.

Nosrati, M. (2011). Python: An appropriate language for real world programming. *World Applied Programming*, *1*(2), 110–117. www.waprogramming.com

Qamal, M., & Fuadi, W. (2022). *ANALISIS SENTIMEN TOKO ONLINE MENGGUNAKAN ALGORITMA NAIVE BAYES CLASSIFIER*.

Que, Valentino & Iriani, Ade & Purnomo, Hindriyanto. (2020). *Analisis Sentimen Transportasi Online Menggunakan Support Vector Machine Berbasis Particle Swarm Optimization*. Jurnal Nasional Teknik Elektro dan Teknologi Informasi. 9. 162-170. 10.22146/jnteti.v9i2.102.

Rizkia, S., Budi, E., Si, S. S., Puspandari, D., & Pd, M. (2023). *Analisis Sentimen Kepuasan Pelanggan Terhadap Internet Provider Indihome di Twitter Menggunakan Metode Decision Tree dan Pembobotan TF-IDF*.

Salman, A. G. (2020), "Konsep Decision Tree & Random Forest," School of Computer Science, https://socs.binus.ac.id/2020/05/26/konsep-decision-tree-random-forest/ (accessed Sep. 30, 2023).

Saputra, A., Subing, M., & Pratama, R. (2023). Perbandingan Metode Naïve Bayes Classifier Dan Support Vector Machine Untuk Analisis Sentimen Pengguna Twitter Mengenai Piala Dunia Fifa 2022 COMPARISON OF NAIVE BAYES TAXONOMY AND SUPPORT VECTOR MACHINES FOR ANALYZING TWITTER USER SENTIMENT ON FIFA WORLD CUP 2022. *TEKNOMATIKA*, *13*(01).

Vijay Gaikwad, S. (2014). Text Mining Methods and Techniques. In *International Journal of Computer Applications* (Vol. 85, Issue 17).