Jurnal Teknologi Sistem Informasi dan Aplikasi **ISSN: 2654-3788**
Penerbit: Program Studi Teknik Informatika Universitas Pamulang **e-ISSN: 2654-4229**
Vol. 6, No. 4, Oktober 2023 (661-670) **DOI: 10.32493/jtsi.v6i3.33814**

Penerapan Algortitma C4.5 untuk Klasifikasi Sentimen Masyarakat terhadap #RUUKUHP pada Twitter

Imam Vusuvangat[1], Siska Kurnia Gusti[2], Fadhilah syafira[3], Novriyanto[4], Fitri insani[5]

Teknik Informatika, Universitas Islam Negeri Sultan Syarif Kasim, Panam, Jl. HR. Soebrantas No.Km. 15, RW.15, Simpang Baru, Kota Pekanbaru, Riau 28293
e-mail: [1]11950115091@students.uin-suska.ac.id, [2]siskakurniagusti@uin-suska.ac.id,
[3]fadhila.syafira@uin-suska.ac.id, [4]novriyanto@enreach.or.id, [5]fitri.insani@uin-suska.ac.id

Submitted Date: August 31st, 2023 Reviewed Date: September 18th, 2023
Revised Date: October 02nd, 2023 Accepted Date: October 04th, 2023

Abstract

Social media, especially Twitter, has developed into an important tool for people to share their opinions and feelings widely. Users often use hashtags to share messages related to certain topics. Some of the issues that lead to the need for sentiment analysis of the Draft Criminal Code are social impact, Public disapproval, Potential legal uncertainty, Potential abuse, Support and criticism. By conducting a sentiment analysis of the draft Penal Code, the government and policymakers can better understand the views of the public, identify possible problems and address them, and make necessary improvements or clarifications to the draft law. This can help ensure that the draft Penal Code has greater public support and adheres to good legal principles. The classification of public responses to this hashtag provides a significant snapshot of public attitudes and perspectives. This study aims to classify public sentiment towards the RUUKUHP hashtag on the Twitter platform using the C4.5 algorithm. This study uses a collection of tweets with the hashtag RUUKUHP which are manually categorized into two and three sentiment categories, namely positive, negative and positive, negative and neutral. In this study, data preprocessing is carried out before training the model which includes removing links, special characters, removing stopwords, and word tokenization. Furthermore, this research uses text representation methods such as TF-IDF to extract features from the tweet text and convert them into numerical vectors used by the C4.5 algorithm. After training the classification model using the C4.5 algorithm with the classified dataset, it evaluates the performance of the model with the metrics of accuracy, recall, precision, and F1 score. Experimental results using 2 categories of Negative and Positive show that the model applied with the C4.5 algorithm achieved an accuracy of 96.6% with a recall of 96.6%, a percision of 97.1% and an F1 score of 96.8. And experiments using 3 categories of Negative, Positive and Neutral achieved an accuracy of 67%, a recall of 67%, a precision of 65%, and an F1 score of 66%. Thus it can be concluded that the results of the RUUKUHP hashtag sentiment classification with 2 class predictions are more relevant than 3 sentiment class predictions with a value reaching 96.6%.

Keywords: Twitter; Community Sentiment; #RUUKUHP; C4.5 Algorithm; Sentiment classification

Abstrak

Media sosial terutama Twitter telah berkembang menjadi alat penting bagi masyarakat untuk berbagi opini dan perasaan secara luas. Pengguna sering menggunakan hashtag untuk membagi pesan yang berkaitan dengan topik tertentu. Beberapa masalah yang menyebab kan perlunya analisis sentimen RUU KUHP adalah dampak sosial, Ketidaksetujuan masyarakat, Potensi Ketidak pastina hukum, Potensi penyalahgunaan, Dukungan dan kritik. Dengan melakukan analisis sentimen terhadap rancangan undang-undang KUHP, pemerintah dan pembuat kebijakan dapat lebih memahami pandangan masyarakat, mengidentifikasi permasalahan yang mungkin timbul dan mengatasinya, serta melakukan perbaikan atau klarifikasi yang diperlukan terhadap naskah rancangan undang-undang tersebut. Hal ini dapat membantu

memastikan bahwa RUU KUHP yang disahkan mendapat dukungan publik yang lebih besar dan mematuhi prinsip-prinsip hukum yang baik. Klasifikasi tanggapan masyarakat terhadap hashtag ini memberikan gambaran yang signifikan tentang sikap dan perspektif publik. Penelitian ini bertujuan untuk mengklasifikasikan sentimen masyarakat terhadap hashtag RUUKUHP di platform *Twitter* dengan menggunakan Algortima C4.5. Dalam Penelitian ini menggunakan kumpulan tweet dengan hashtag RUUKUHP yang dikategorikan secara manual menjadi dua dan tiga kategori sentimen yaitu positif, negatif dan positif, negatif dan netral. Pada penelitian ini dilakukan preprocessing data sebelum melatih model yang mencakup penghapusan tautan, karakter khusus, penghilangan stopwords, dan tokenisasi kata. Selanjutnya penelitian ini menggunakan metode representasi teks seperti TF-IDF untuk mengekstrak fitur dari teks tweet dan mengubahnya menjadi vektor numerik yang digunakan oleh algoritma C4.5. Setelah melatih model klasifikasi menggunakan algoritma C4.5 dengan dataset yang telah diklasifikasikan, dalam mengevaluasi kinerja model dengan metrik akurasi, recall, presisi, dan skor F1. Hasil eksperimen menggunakan 2 kategori Negatif dan Positif menunjukkan bahawa model yang diterapkan dengan algoritma C4.5 mencapai akurasi sebesar 96,6% dengan recall 96,6%,percision 97,1% dan skor F1 96,8.Dan eksperimen yang menggunakan 3 kategori Negatif,Positif dan Netral mencapai akurasi sebesar 67%, recall sebesar 67%, presisi sebesar 65%, dan skor F1 sebesar 66%.Dengan demikian dapat disimpulkan bahawa hasil dari klasifikasi sentimen hashtag RUUKUHP dengan 2 prediksi kelas yang lebih relevan dibandingkan 3 prediksi kelas sentiment dengan nilai mencapai 96,6%.

Kata Kunci:*Twitter; Sentimen Masyarat; #RUUKUHP; Algoritma C4.5; klasifikasi Sentimen*

1 Pendahuluan

Media sosial telah berkembang menjadi platform penting dalam era digital yang terus berkembang untuk menyampaikan pendapat dan ekspresi pengguna tentang berbagai topik, seperti ulasan dan diskusi tentang ponsel pintar. Dalam hal ini, Hashtag RUUKUHP digunakan untuk mengumpulkan konten yang berkaitan dengan ulasan, perbandingan, dan pertanyaan tentang ponsel pintar(Seno & Wibowo, 2019). Pada hashtag RUUKUHP ini menggunakan klasifikasi sentimen.

Klasifikasi sentimen terhadap hashtag RUUKUHP bertujuan untuk mengevaluasi sentimen atau sikap pengguna terhadap ponsel pintar yang dibahas dalam konten dengan hashtag tersebut. Algoritma C4.5 adalah salah satu cara untuk melakukan klasifikasi sentimen. Algoritma C4.5 ini adalah Algoritma pembelajaran mesin yang digunakan dalam pembuatan model keputusan berbasis pohon. Algoritma ini membangun pohon keputusan berdasarkan atribut-atribut yang relevan dalam dataset yang berisi teks atau atribut teks terkait ulasan ponsel pintar(Mardi, 2017) . Pohon keputusan ini dapat digunakan untuk mengklasifikasikan ulasan ulasan tersebut ke dalam kategori sentimen positif, negatif, atau netral. Klasifikasi sentimen adalah salah satu cabang dari text mining. Klasifikasi emosi dapat menjadi bagian penting dalam menilai topik masalah.

Tujuan utama dari klasifikasi perasaan adalah untuk mengeksplorasi polaritas emosi positif, negatif, dan netral. Salah satu pengklasifikasian sentimen dapat diperoleh melalui kicauan di Twitter(Suryono et al., 2018). Dalam artikel ini, tweet terkait kata kunci dicari berdasarkan hashtag dan dikumpulkan menggunakan alat termasuk Twitter API. Data yang diperoleh selama pengumpulan akan diolah menggunakan pemrosesan bahasa alami yang dijalankan dengan bahasa pemrograman Python. Data tersebut kemudian akan dirangking berdasarkan sentimen menggunakan algoritma C4.5 untuk melihat hasil sentimen.

Dalam klasifikasi sentimen menggunakan algoritma C4.5 terhadap hashtag RUUKUHP, langkah-langkah yang umum dilakukan meliputi pra-pemrosesan data, seperti pembersihan dan normalisasi teks, ekstraksi fitur, dan pembentukan dataset training(Anggraini & Utami, 2021). Kemudian algoritma C4.5 diterapkan untuk membangun model keputusan berdasarkan dataset tersebut. Model ini kemudian dapat digunakan untuk mengklasifikasikan ulasan ponsel pintar yang menggunakan hashtag RUUKUHP ke dalam kategori sentimen yang sesuai.

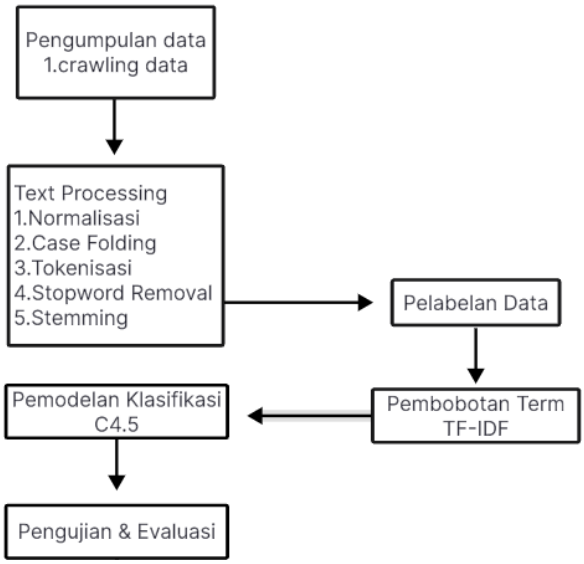
Penelitian ini bertujuan untuk melakukan klasifikasi sentimen menggunakan algoritma C4.5

terhadap hashtag RUUKUHP dengan dua kali pengujian yang dapat dikategorikan 2 kelas sentimen postif, negatif dan 3 kelas sentimen positif, negatif, dan netral.

Dalam penelitian ini penulis menggunakan kategori positif dan negatif karena hasil penelitian menunjukkan bahwa dengan menggunakan positif dan negatif dapat memperloleh akurasi yang tinggi pada metode yang digunakan.

2 Metode Penelitian

Pada tahap ini adalah langkah langkah yang akan dilakukan dengan metode algortima c4.5 dapat diperhatikan pada gambar 1 berikut



Gambar 1.Alur Penelitian

Penjelasan Gambar 1 adalah sebagai berikut:

1. Pengumpulan Data

   Pada tahap ini, peneliti mengumpulkan data menggunakan Crawling data Twitter menggunakan library Python. Tahapan ini sangat penting karena dapat memengaruhi hasil penelitian, sehingga data harus dikumpulkan dengan benar. Data yang digunakan dalam penelitian ini diperoleh dari hastag #RUU KUHP(Mailo & Lazuardi, 2019). Data yang terkumpul sebanyak 6069 data.
2. Analysis sentiment

   Adalah menggunakan pemrosesan bahasa alami, analisis teks, linguistik komputasi, dan biometrik untuk mengidentifikasi, mengekstrak, mengukur, dan mempelajari keadaan afektif dan informasi subjektif(Muzaki et al., 2023).Atau bias disebut juga dengan proses memahami dan mengelompokan emosi(positif,negative,neutral)(Mubarok, 2021) pada tulisan dengan menggunakan analisis teks.
3. Text Mining

   adalah proses ekstraksi, pemrosesan, dan analisis teks untuk mengidentifikasi pola, tren, dan wawasan yang terkandung dalam teks yang tidak terstruktur. Hal ini dilakukan dengan menggunakan metode seperti pemrosesan bahasa alami (NLP)(Herianto, 2019), pembelajaran mesin, dan penambangan asosiasi. Tujuannya adalah mengubah Teks tidak terstruktur dapat dianggap sebagai data terstruktur yang dapat dianalisis untuk lebih memahami informasi yang terkandung dalam teks. Text Mining membuka peluang untuk mengeksplorasi dan memanfaatkan potensi informasi yang terkandung dalam teks dengan lebih efisien(A. Hermawan et al., 2023).
4. Text Analysis

   Analisis teks adalah teknik dan prosedur yang digunakan untuk mengumpulkan informasi dan pengetahuan berkualitas dari data teks(Mailoa, 2021). Ini biasanya digunakan untuk memodelkan dan mengekstraksi informasi dari dokumen kunci untuk analisis bisnis dan tujuan lainnya.
5. Text Preprocessing

   Preprocessing teks dilakukan untuk menghilangkan noise data seperti simbol, tanda baca, singkatan, kesalahan ketik, kata-kata yang tidak masuk akal dan lain-lain. Pemrosesan awal teks biasanya melibatkan beberapa langkah, seperti pembersihan, case folding, tokenisasi, stemming, dan pemfilteran (L. Hermawan & Bellaniar Ismiati, 2020). Text preprocessing juga dapat digunakan untuk mengatasi noise, seperti normalisasi bahasa. Namun, tahapan stemming tidak dilakukan dalam penelitian ini karena dapat mengurangi akurasi model
   a. Cleaning

      Penanganan untuk menghapus karakter yang dapat mengganggu

pemrosesan lebih lanjut, seperti angka, tanda baca, URL, nama pengguna, dll., prosedur penghapusan pengguna, dll.

b. Case Folding

Proses merubah huruf besar bertujuan untuk mengubah bentuk setiap huruf menjadi format yang sama, yaitu huruf besar atau huruf kecil. .

c. Tokenizing

Proses ini menggunakan token untuk memecah setiap dokumen menjadi kata-kata yang membentuk dokumen tersebut.

d. Normalisasi

Bahasa berarti mengembalikan kata-kata yang tidak baku atau tidak sesuai dengan KBBI, seperti slang (gaul). Misalnya, kata-kata seperti "y" dan "ya" diubah menjadi "iya".

e. Stemming

Selama proses penyaringan atau filter, kata-kata umum atau tidak berarti dihapus, meninggalkan kata-kata yang tersisa bermakna .

6. Filtering

Proses filtering menghilangkan kata-kata yang sering digunakan atau tidak berarti, sehingga kata-kata yang tersisa memiliki arti yang signifikan.

7. Pembobotan TF-IDF

Nilai kata (term) diperoleh melalui pembobotan kata ini. Setelah preprocessing, data harus berbentuk numerik. Untuk melakukan ini, Bobot TF-IDF digunakan untuk menentukan seberapa dekat suatu kata (term) terkait dengan dokumen dengan memberikan bobot pada setiap kata. (Sidiq et al., 2020). TF-IDF menggabungkan dua konsep frekuensi kata muncul dalam dokumen dan invers frekuensi dokumen yang mengandung kata. Menghitung bobot dengan TF-IDF

Untuk menghitung pembobot TF-IDF, rumusnya adalah sebagai berikut:

$$\boldsymbol{Idf} = \log \frac{\boldsymbol{tf}}{\boldsymbol{log}} \left(\frac{\boldsymbol{N}}{\boldsymbol{df}}\right) \ldots (1)$$

$$\boldsymbol{W} = \boldsymbol{tf} * \boldsymbol{idf} \ldots\ldots\ldots (2)$$

Keterangan:
d = dokumen
t = kata pada dokumen
W = bobot dokumen ke-d dengan kata ke-t
Tf = banyaknya kemunculan term(kata) dalam dokumen
Ft,d = frekuensi kata pada d
IDF = Inversed Document Frequency.
Dft = banyak kata yang mengandung kata i.
N = jumlah seluruh dokumen

Nilai perbandingan dibagi antara jumlah kata yang muncul dalam dokumen dibagi dengan jumlah kata yang muncul dalam dokumen sehingga jumlah kata yang muncul dalam dokumen sama dengan satu disebut sistem klasifikasi. Alternatifnya, satu dokumen frekuensi pengembalian (idf) atau kombinasi TF-IDF dapat digunakan untuk setiap kata.

8. Decision Tree

Pohon keputusan adalah struktur yang digunakan untuk membagi kumpulan data besar menjadi kumpulan rekaman yang lebih kecil sehingga serangkaian aturan keputusan dapat diterapkan. Decision Tree merupakan cara untuk mengubah kejadian besar menjadi pohon keputusan yang memprediksi aturan, salah satu algoritma yang digunakan dalam pelatihan pohon keputusan adalah Algoritma C4.5 (Somantri & Dairoh, 2019). Algoritma ID3 digunakan untuk menjalankan proses, dengan input sampel instruksi, label instruksi, dan atribut. Algoritma C4.5 merupakan evolusi dari ID3 (Iterative Dichotomizer 3). Beberapa pengembangan yang dilakukan pada C4.5 antara lain kemampuan untuk menangani missing value, memperbaiki pemotongan, dan memperbaiki data berulang (Puspita & Widodo, 2021). Nilai gain tertinggi dari atribut harus digunakan untuk menentukan atribut. Persamaan mengandung rumus berikut.

Rumus Entropy

$$Entropy(S) = \sum_{t=1}^{n} -pi * log2pi \ldots (3)$$

Keterangan :
S = Himpunan Kasus
n = Jumlah Partisi S
pi = Proporsi dari Si terhadap S

Rumus Gain

$$\text{Gain (S,A)} = \text{Entropy (S)} = \sum_{i=1}^{n} \frac{|Si|}{|S|} * Entropy\ (Si) \ldots (4)$$

Keterangan
S = Himpunan Kasus
A = Atribut
n = Jumlah Pratisi A
|*Si*|=Jumlah kasus pada pratisi ke i
| S | = Jumlah kasus dalam S

9. Klasifikasi Data Mining

adalah proses menempatkan objek atau contoh ke dalam kategori atau kelas yang sudah ditentukan berdasarkan pola dan fitur yang ada dalam data. Tujuannya adalah untuk memprediksi kelas atau label dari objek baru berdasarkan data latih yang sudah diketahui. Algoritma klasifikasi, seperti Naive Bayes, Decision Tree, k-Nearest Neighbors, Support Vector Machines, dan Random Forests (Permana et al., 2021), digunakan untuk membangun model klasifikasi dan mengklasifikasikan objek baru ke dalam kelas yang tepat. Klasifikasi data mining memiliki berbagai aplikasi di berbagai bidang, dan dapat membantu organisasi membuat keputusan yang lebih baik dan memahami data dengan lebih baik(Anggada Maulana, 2018).

a. Algoritma C4.5

Merupakan teknik untuk mengklasifikasikan data menggunakan model pohon keputusan. Algoritma ini terdiri dari beberapa langkah. Yang pertama adalah mengatur atribut sebagai akar, cabang untuk setiap nilai dan membagi kasus menjadi cabang (Siahaan et al., 2020). Algoritma ini memiliki beberapa kelebihan selain digunakan untuk membuat pohon keputusan. Penanganan data numerik dan diskrit adalah salah satu kelebihannya. Algoritma ini terutama digunakan untuk membuat pohon keputusan dengan memilih atribut dengan prioritas tertinggi (juga dikenal sebagai memiliki nilai gain tertinggi) berdasarkan nilai entropy mining dan pengambilan informasi(Haqmanullah Pambudi et al., 2018). Algoritma ini merupakan perangkat lunak yang dapat digunakan secara gratis.

b. Confussion Matrix

Confussion matrix adalah tabel yang menunjukkan distribusi terstruktur dari data uji benar dan salah (Wulandari et al., 2022). Tujuan utama dengan confussion matrix adalah untuk memberikan perbandingan hasil klasifikasi yang dilakukan oleh sistem dengan hasil klasifikasi yang sebenarnya. Kedua algoritma ini harus digunakan untuk memprediksi persediaan yang didukung oleh confussion matrix (Fitriani et al., 2020), yang menentukan persentase akurasi, presisi, dan recall.

3 Hasil dan Pembahasan

3.1 Hasil Pengumpulan Data

Data yang diperoleh tersebut didapatkan melalui proses Crawling di media sosial Twitter dengan menggunakan API Twitter. Jumlah data yang didapatkan menghasilkan 6069 tweet.Data tersebut di dapatkan dari 01-08-2022 sampai dengan 13-12-2022

3.2 Preproccesing Text

Pada tahap preprocessing, sebelum menggunakan data pada tweet, data bersih harus didapatkan dan diberi label. Langkah-langkah preprocessing sebagai berikut.

a. Case folding

Pada tahap ini Prosedur mengubah semua data yang ada ditext dari huruf besar menjadi kecil atau sebaliknya. Dapat diperhatikan pada Tabel 1.

Tabel 1.Case Folding

Data *Input*	Data *Output*
b’Dispar Kepri Jamin RUU KUHP yang Baru Tak Ganggu Pariwisata	b’dispar kepri jamin ruu kuhp yang baru tak ganggu pariwisata

b. Cleaning

Selanjutnya cleaning yaitu untuk membersihkan data dari karakter yang tidak diperlukan seperti tanda baca dan beberapa karakter lainnya.data yang cleaning di hasil kan menjadi 3286 data tweet yang sebelum

ISSN: 2654-3788
e-ISSN: 2654-4229
DOI: 10.32493/jtsi.v6i3.33814

nya data 6069 data tweet Dapat diperhatikan pada Tabel 2.

Tabel 2.Cleaning

Data *Input*	Data *Output*
b'dispar kepri jamin ruu kuhp yang baru tak ganggu pariwisata	dispar kepri jamin ruu kuhp yang baru tak ganggu pariwisata

c. Tokenizing

Selanjutnya tokenizing yaitu tahap ini memotong setiap kalimat menjadi kata. Dapat diperhatikan pada Tabel 3.

Tabel 3. Tokenizing

Data *Input*	Data *Output*
dispar kepri jamin ruu kuhp yang baru tak ganggu pariwisata	['dispar', 'kepri', 'jamin', 'ruu', 'kuhp', 'yang', 'baru', 'tak', 'ganggu', 'pariwisata']

d. Stopword Removal

Sesudah itu stopword removal yaitu melepaskan kata yang tidak penting. Dapat diperhatikan pada Tabel 4

Tabel 4. Stopword Removal

Data *Input*	Data *Output*
['dispar', 'kepri', 'jamin', 'ruu', 'kuhp', 'yang', 'baru', 'tak', 'ganggu', 'pariwisata']	['dispar', 'kepri', 'jamin', 'ruu', 'kuhp', 'ganggu', 'pariwisata']

e. Stemming

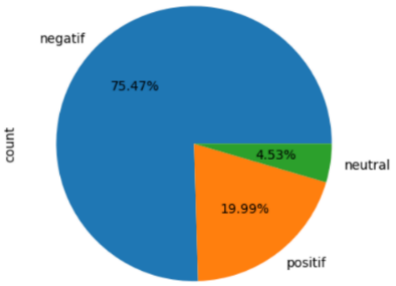
Selanjutnya yang terakhir stemming adalah pemotongan kata-kata imbuhan menjadi kata dasar. Dapat diperhatikan pada Tabel 5

Tabel 5.Stemming

Data *Input*	Data *Output*
['dispar', 'kepri', 'jamin', 'ruu', 'kuhp', 'ganggu', 'pariwisata']	dispar kepri jamin ruu kuhp ganggu pariwisata

3.3 Pelabelan Data

Pelabelan data ini sudah melewati tahap preprocessing dengan jumlah data yang di hasilkan sebanyak 3286 data tweet dari yang sebelum nya 6069 data tweet yang didapat. Pelabelan data sentimen yang di label kan oleh ahli bahsa menjadi 3 bagian, yaitu positif, negatif dan netral. Pelabelan data sentimen positif menghasilkan 657 tweet, pelabelan data sentimen negatif menghasilkan 2.480 tweet, dan pelabelan data sentimen netral menghasilkan 169 tweet, Dapat dilihat pada gambar 2.



Gambar 2.Presentase #RUUKUHP

3.4 Pembobotan TF-IDF

Kemudian langkah pembobotan TF-IDF menghitung kata-kata yang muncul. Nilai bobot pada dokumen TF-IDF (Term Frekuensi Inverted Document Frekuensi) diperoleh berdasarkan frekuensi kemunculan kata tersebut. Hasil pembobotan TF-IDF dapat diperhatikan pada tabel 6

Tabel 6.TF-IDF

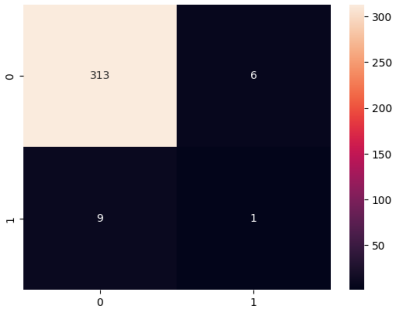
TERM	TF-IDF		
	D1	D2	D3
dispar	0,477	0	0
kepri	0,477	0	0
jamin	0,477	0	0
ruu	0,477	0	0
kuhp	-0,221	-0,665	-0,221
……	……	……	……
dulunya	0	0	0,477
buatan	0	0,352	0
kolonia	0	0,477	0
belanda	0	0,477	0
……	……	……	……
pasal	0	0,477	0
karet	0	0,477	0
antidemo	0	0,477	0
menyeret	0	0,477	0
……	……	……	……

3.5 Klasifikasi C4.5

Setelah itu, data berita yang telah melalui text preprocessing dan word weights (tf-idf) akan dilakukan proses klasifikasi menggunakan metode C4.5 untuk membuat model yang mampu mengklasifikasikan data baru berdasarkan subjek tanpa pelabelan manual. Pada tahap ini dilakukan proses pembelajaran dan pelatihan. Pembelajaran dilakukan dengan melatih model menggunakan kumpulan data berlabel sedangkan pembelajaran dilakukan dengan menggunakan data tidak berlabel. Hasil klasifikasi C4.5 pada penelitian ini cukup baik dengan akurasi yang tinggi, algoritma ini dinilai cocok untuk mengklasifikasikan data berdasarkan topik.

3.6 Pengujian Confusion Matrix dengan 2 kelas sentimen positif dan negative

a. Hasil 90% train dan 10% test merupakan hasil pengujian matriks konfusi dengan menggunakan 90% train dan 10% test pada Gambar 3 di bawah ini.

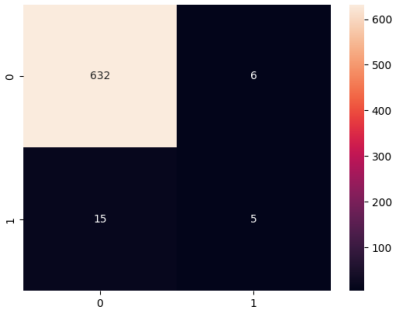


Gambar 3. Perbandingan 90:10%

$$Accuracy = \frac{TP+TN}{JulahData} x100\%$$

$$Accuracy = \frac{313+1}{312+6+9+1} x100\%$$
$$= \frac{314}{329} x100\%$$
$$= 95{,}44/\%$$

b. Hasil 80% train dan 20% test merupakan hasil pengujian matriks konfusi dengan menggunakan 80% train dan 20% test pada Gambar 4 di bawah ini.

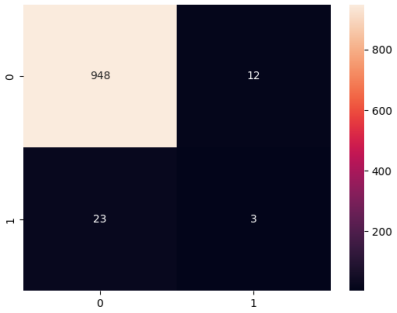


Gambar 4. Perbandingan 80:20%

$$Accuracy = \frac{TP+TN}{JulahData} x100\%$$

$$Accuracy = \frac{632+5}{632+5+15+5} x100\%$$
$$= \frac{637}{658} x100$$
$$= 96{,}80\%$$

c. Hasil 70% train dan 30% test merupakan hasil pengujian matriks konfusi dengan menggunakan 70% train dan 30% test pada Gambar 5 di bawah ini.



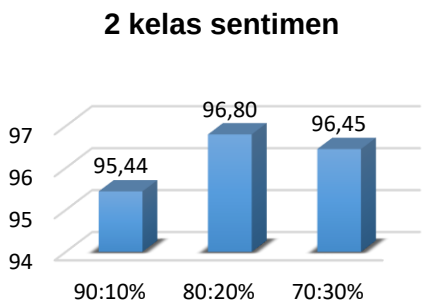
Gambar 5. Perbandingan 70:30%

$$Accuracy = \frac{TP+TN}{JulahData} x100\%$$

$$Accuracy = \frac{948+3}{948+12+23+3} x100\%$$
$$= \frac{951}{986} x100\%$$
$$= 96{,}45\%$$

d. Hasil Grafik 2 Kelas Sentimen
Berdasarkan hasil uji dari 2 kelas sentimen dapat dilihat dari grafik gambar 6 di bawah ini.

ISSN: 2654-3788
e-ISSN: 2654-4229
DOI: 10.32493/jtsi.v6i3.33814

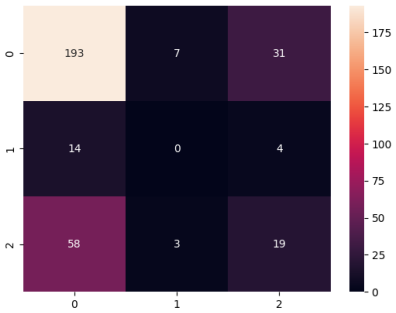


Gambar 6. Grafik Presentasi

Dari gambar grafik di atas dapat di lihat melakukan pengujian data sebanyak 3 kali dengan pengujian 2 kelas sentimen dari hasil akurasi yang tertinggi mendapat kan hasil akurasi sebesar 96,80% dengan data latih 80% serta 20% data uji

3.7 Pengujian Confusion Matrix dengan 3 kelas sentimen positif negatif dan netral

a. Hasil 90% train dan 10% test merupakan hasil pengujian matriks konfusi dengan menggunakan 90% train dan 10% test pada Gambar 7 di bawah ini.



Gambar 7. Perbandingan 90:10%

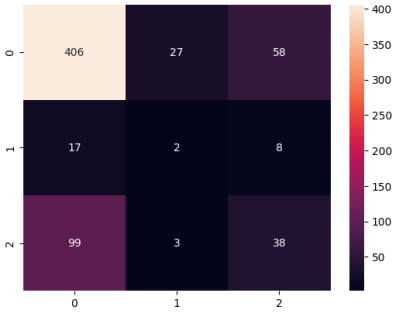
$$Accuracy = \frac{TN + TN + TP}{JulahData} x100\%$$

$$Accuracy = \frac{193+0+19}{193+7+31+14+0+4+58+3+19} x100\%$$

$$= \frac{212}{329} x100\%$$

$$= 64{,}43\%$$

b. Hasil 80% train dan 20% test merupakan hasil pengujian matriks konfusi dengan menggunakan 80% train dan 20% test pada Gambar 8 di bawah ini.



Gambar 8. Perbandingan 80:20%

$$Accuracy = \frac{TN + TN + TP}{JulahData} x100\%$$

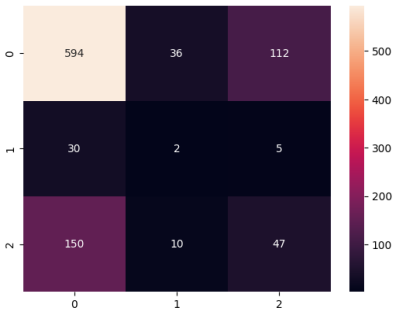
$$Accuracy = \frac{406+2+38}{406+27+58+17+2+8+99+3+38} x100\%$$

$$= \frac{446}{658} x100\%$$

$$= 67{,}78\%$$

c. Hasil 70% train dan 30% test merupakan hasil pengujian matriks konfusi dengan menggunakan 70% train dan 30% test pada Gambar 9 di bawah ini.

d.



Gambar 9. Perbandingan 70:30%

$$Accuracy = \frac{TN + TN + TP}{JulahData} x100\%$$

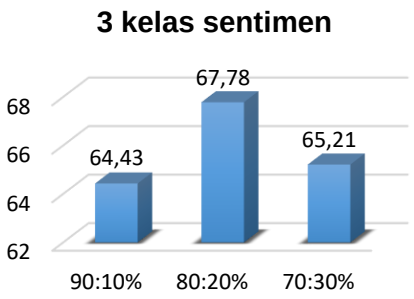
$$Accuracy = \frac{594+2+47}{594+36+112+30+2+5+150+10+47} x100\%$$

$$= \frac{643}{986} x100\%$$

$$= 65{,}21\ \%$$

e. Hasil Grafik 3 kelas Sentimen
Berdasarkan hasil uji dari 2 kelas sentimen dapat dilihat dari grafik gambar 10 di bawah ini.

ISSN: 2654-3788
e-ISSN: 2654-4229
DOI: 10.32493/jtsi.v6i3.33814



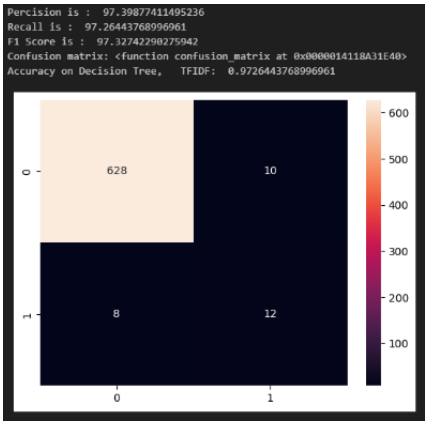
Gambar 10. Grafik Presentase

Dari gambar grafik di atas dapat di lihat melakukan pengujian data sebanyak 3 kali dengan pengujian 3 kelas sentimen dari hasil akurasi yang tertinggi mendapat kan hasil akurasi sebesar 67,78% dengan data latih 80% serta 20% data uji

3.8 Evaluasi

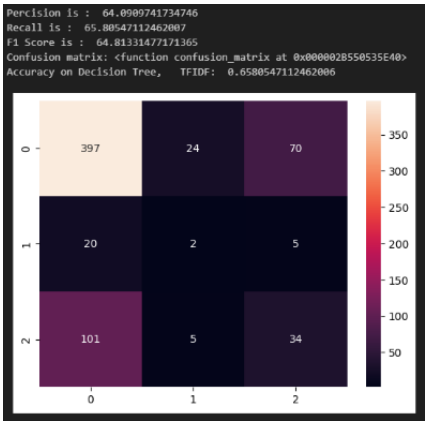
Setelah proses klasifikasi dan analisis sentimen, Langkah berikutnya adalah pengujian. Dilakukan proses klasifikasi dan evaluasi menggunakan model klasifikasi teks Multinomial Decision Tree. Hasil evaluasi digunakan untuk mengevaluasi akurasi model Decision Tree dan hasil menggunakan Confusion Matrix

1. Menggunakan 2 kelas sentimen mendapatkan hasil evaluasi sebagai uji accuracy Decession tree sebesar 97,26% precission sebesar 97%, recall sebesar 97%, dan nilai f1-score sebesar 97% dapat diperhatikan pada Gambar 9 di bawah ini.



Gambar 11.Evaluasi 2 kelas

2. Menggunakan 3 kelas sentimen mendapatkan hasil evaluasi sebagai uji accuracy Decession tree sebesar 68,88% precission sebesar 64%, recall sebesar 65%, dan nilai f1-score sebesar 64% dapat diperhatikan pada gambar 10 di bawah ini.



Gambar 12. Hasil evaluasi 3 kelas

4 Kesimpulan

Berdasarkan dari hasil pengujian yang dilakukan pada tahap penelitian ini dapat diambil beberapa kesimpulan sebagi berikut:

1. Algoritma C4.5 dapat diterapkan untuk menentukan peringkat sentimen publik untuk tagar #RUUKUHP berdasarkan komentar twitter
2. Berdasarkan hasil pengujian yang tertinggi di katagori 2 kelas sentimen
3. Berdasarkan hasil pengujian data yang terbesar dan tertinggi dapat ditemukan pada pengujian dengan 80% data latih dan 20% data uji pada kalsifikasi menggunakan 2 kelas Negatif dan Positif dengan akurasi 96,8% Percision 96.0% recall 96% dan f1 scsore 96,3%
4. Berdasarkan hasil pengujian yang menggunakan 3 kelas sentimen mendapatkan hasil akurasi 68,88% precission sebesar 64%, recall sebesar 65%, dan nilai f1-score sebesar 64%
5. Berdasarkan klasifikasi yang telah dilakukan hasil yang tertinggi dan yang relevan adalah menggunakan 2 kelas sentimen Negatif dan Positif.

References

Anggada Maulana. (2018). Konsep Dasar Data Mining. *Konsep Data Mining*, *1*, 1–16.

Anggraini, W. P., & Utami, M. S. (2021). Klasifikasi Sentimen Masyarakat Terhadap Kebijakan Kartu Pekerja Di Indonesia. *Faktor Exacta*, *13*(4), 255. https://doi.org/10.30998/faktorexacta.v13i4.796

4

Fitriani, E., Aryanti, R., Saepudin, A., & Ardiansyah, D. (2020). Penerapan Algoritma C4.5 Untuk Klasifikasi Penempatan Tenaga Marketing. *Paradigma - Jurnal Komputer Dan Informatika*, *22*(1), 72–78. https://doi.org/10.31294/p.v22i1.6898

Haqmanullah Pambudi, R., Darma Setiawan, B., & Indriati. (2018). Penerapan Algoritma C4.5 Untuk Memprediksi Nilai Kelulusan Siswa Sekolah Menengah Berdasarkan Faktor Eksternal. *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer*, *2*(7), 2637–2643. http://j-ptiik.ub.ac.id

Herianto. (2019). *Penerapan Text-Mining Untuk Mengidentifikasi*. *VIII*(2), 36–44.

Hermawan, A., Jowensen, I., Junaedi, J., & Edy. (2023). Implementasi Text-Mining untuk Analisis Sentimen pada Twitter dengan Algoritma Support Vector Machine. *JST (Jurnal Sains Dan Teknologi)*, *12*(1), 129–137. https://doi.org/10.23887/jstundiksha.v12i1.52358

Hermawan, L., & Bellaniar Ismiati, M. (2020). Pembelajaran Text Preprocessing berbasis Simulator Untuk Mata Kuliah Information Retrieval. *Jurnal Transformatika*, *17*(2), 188. https://doi.org/10.26623/transformatika.v17i2.1705

Mailo, F. F., & Lazuardi, L. (2019). Analisis Sentimen Data Twitter Menggunakan Metode Text Mining Tentang Masalah Obesitas di Indonesia. *Journal of Information Systems for Public Health*, *4*(1), 28–36.

Mailoa, F. F. (2021). Analisis sentimen data twitter menggunakan metode text mining tentang masalah obesitas di indonesia. *Journal of Information Systems for Public Health*, *6*(1), 44. https://doi.org/10.22146/jisph.44455

Mardi, Y. (2017). Data Mining : Klasifikasi Menggunakan Algoritma C4.5. *Edik Informatika*, *2*(2), 213–219. https://doi.org/10.22202/ei.2016.v2i2.1465

Mubarok, R. (2021). Analisis Sentimen Pengguna Twitter Terhadap Kebijakan Pemberlakuan Pembatasan Sosial Berskala Besar (Psbb) Dengan Metode …. *Jurnal Siliwangi Seri Sains Dan Teknologi*, *7*(1), 19–24. http://jurnal.unsil.ac.id/index.php/jssainstek/article/view/3726

Muzaki, H., Marcos, H., Letjend, J., Soemarto, P., Utara, K. P., & Banyumas, K. (2023). *Analisis Sentimen RUU KUHP di Medsos Twitter dengan Metode Naïve Bayes*. *7*(1), 1–4.

Permana, A. P., Ainiyah, K., & Holle, K. F. H. (2021). Analisis Perbandingan Algoritma Decision Tree, kNN, dan Naive Bayes untuk Prediksi Kesuksesan Start-up. *JISKA (Jurnal Informatika Sunan Kalijaga)*, *6*(3), 178–188. https://doi.org/10.14421/jiska.2021.6.3.178-188

Puspita, R., & Widodo, A. (2021). Perbandingan Metode KNN, Decision Tree, dan Naïve Bayes Terhadap Analisis Sentimen Pengguna Layanan BPJS. *Jurnal Informatika Universitas Pamulang*, *5*(4), 646. https://doi.org/10.32493/informatika.v5i4.7622

Seno, D. W., & Wibowo, A. (2019). Analisis Sentimen Data Twitter Tentang Pasangan Capres-Cawapres Pemilu 2019 Dengan Metode Lexicon Based Dan Support Vector Machine. *Jurnal Ilmiah FIFO*, *11*(2), 144. https://doi.org/10.22441/fifo.2019.v11i2.004

Siahaan, S. W., Sianipar, K. D. R., R.H Zer, P. P. P. A. N. . F. I., & Hartama, D. (2020). Penerapan Algoritma C4.5 Dalam Meningkatkan Kemampuan Bahasa Inggris Pada Mahasiswa. *Petir*, *13*(2), 229–239. https://doi.org/10.33322/petir.v13i2.1029

Sidiq, R. P., Dermawan, B. A., & Umaidah, Y. (2020). Sentimen Analisis Komentar Toxic pada Grup Facebook Game Online Menggunakan Klasifikasi Naïve Bayes. *Jurnal Informatika Universitas Pamulang*, *5*(3), 356. https://doi.org/10.32493/informatika.v5i3.6571

Somantri, O., & Dairoh, D. (2019). Analisis Sentimen Penilaian Tempat Tujuan Wisata Kota Tegal Berbasis Text Mining. *Jurnal Edukasi Dan Penelitian Informatika (JEPIN)*, *5*(2), 191. https://doi.org/10.26418/jp.v5i2.32661

Suryono, S., Utami, E., & Luthfi, E. T. (2018). Klasifikasi Sentimen Pada Twitter Dengan Naive Bayes Classifier. *Angkasa: Jurnal Ilmiah Bidang Teknologi*, *10*(1), 89. https://doi.org/10.28989/angkasa.v10i1.218

Wulandari, Y., Haerani, E., Gusti, S. K., & Ramadhani, S. (2022). Klasifikasi Berita Menggunakan Algoritma C4.5. *Jurnal Nasional Komputasi Dan Teknologi Informasi (JNKTI)*, *5*(2), 279–289. https://doi.org/10.32672/jnkti.v5i2.4194