Jurnal_Nas_Terakreditasi_q.pdf

by

Submission date: 19-Jun-2023 10:27PM (UTC+0700)
Submission ID: 2119114123
File name: Jurnal_Nas_Terakreditasi_q.pdf (561.89K)
Word count: 4878
Character count: 27334

Jurnal Nasional Komputasi dan Teknologi Informasi
Vol. 5 No. 2, April 2022
P-ISSN 2620-8342
E-ISSN 2621-3052

Klasifikasi Berita Menggunakan Algoritma C4.5

Yayuk Wulandari[1], Elin Haerani[2], Siska Kurnia gusti[3]
dan Siti Ramadhani[4]
[1234] Teknik Informatika Universitas Islam Negeri Sultan Syarif Kasim Riau
Jl. H.R Soebrantas no.155 KM.18 Simpang Baru, Pekanbaru 28293
Corresponding author's e-mail: 11850122344@students.uin-suska.ac.id[1] , elin.haerani@uin-suska.ac.id[2] , siskakurniagusti@uin-suska.ac.id[3] , siti.ramadhani@uin-suska.ac.id[4]

Abstrak - Perkembangan zaman mengalami kemajuan yang sangat pesat, saat ini penyebaran berita yang paling populer adalah melalui internet. Berita yang disajikan di situs berita online biasanya hanya dalam kategori umum, sehingga ketika pembaca ingin mendapatkan kategori berita yang lebih spesifik harus dilakukan secara manual yang tentunya menjadi kegiatan yang cukup merepotkan. Hal ini juga dialami oleh Badan Pusat Statistik Provinsi Riau yang kesulitan dalam mencari dan mengklasifikasikan berita tentang Provinsi Riau. Dalam hal ini penerapan klasifikasi otomatis dirasa sangat diperlukan. Penelitian ini menggunakan metode klasifikasi C4.5 dengan 510 data berita yang akan diklasifikasikan menjadi 3 kategori yaitu demokrasi, kemiskinan dan ketenagakerjaan. Proses klasifikasi berita dalam penelitian ini meliputi: pengumpulan data, pelabelan manual, teks preprocessing, pembobotan kata, dan metode klasifikasi C4.5. Berdasarkan penelitian yang dilakukan, hasil uji akurasi adalah 84% dengan menggunakan pembagian data 90%:10%. Dapat disimpulkan bahwa metode C4.5 cocok digunakan dalam klasifikasi berita.
Kata kunci: Badan Pusat Statistik, Berita, C4.5, Klasifikasi.

Abstract - *The development of the times has progressed very rapidly, currently the most popular spread of news is through the internet. The news presented on online news sites is usually only in general categories, so when readers want to get a more specific category of news, it must be done manually, which of course will be a bit of a hassle. This is also experienced by the social sector of the Badan Pusat Statistik of Riau, which has difficulty finding and classifying news about Riau Province. In this case the application of automatic classification is felt to be very necessary. This study uses the C4.5 classification method with 510 news data which will be classified into 3 categories, namely democracy, poverty and employment. The news classification process in this study includes: data collection, manual labeling, preprocessing text, word weighting, and C4.5 classification method. Based on the research conducted, the results of the accuracy test were 84% using 90%:10% data sharing. It can be concluded that the C4.5 method is suitable for use in news classification.*
Keywords : Badan Pusat Statistik, C4.5, Classification, News.

1. Pendahuluan

Berita adalah informasi yang menyampaikan suatu keadaan atau peristiwa terkini untuk memberitahu masyarakat luas tentang suatu keadaan. Perkembangan zaman begitu pesat sehingga kini banyak cara penyampaian berita tidak hanya melalui media cetak tetapi juga melalui radio, televisi dan salah satunya yang paling populer yaitu melalui internet. Penyebaran berita melalui internet seperti blog, website dan media sosial memang sangat popular, dikarenakan masyarakat yang sebagian besar adalah pengguna internet lebih mudah untuk mengakses berbagai berita terbaru disana [1]. Dengan tingginya minat masyarakat dalam mengakses berita secara online, membuat situs portal berita harus bekerja lebih keras demi menyediakan informasi yang berkualitas untuk masyarakat.

Pada situs berita online seperti website biasanya sudah terdapat beberapa kategori berita yang telah dikelompokkan, hal ini bertujuan untuk mempermudah pembaca saat hendak mencari berita yang diinginkan secara cepat. Namun pengelompokan berita tersebut masih termasuk kategori umum, jadi jika pembaca ingin mendapatkan kategori berita yang lebih spesifik maka harus dilakukan secara manual dengan memfilter berita dalam kategori tersebut lalu memecahnya menjadi subkategori yang lebih detail. Hal tersebut menjadi kesulitan tersendiri karena diperlukan ketelitian dalam membaca dan menyaring berita yang biasanya memiliki tingkat similaritas yang tinggi, ditambah lagi jumlah berita setiap waktu pastinya terus meningkat dengan cepat.

Proses klasifikasi termasuk ke dalam salah satu bidang *data mining*, yaitu pada *text mining*.Text mining adalah penerapan konsep dan teknik data mining saat melakukan proses analisis teks untuk mendapatkan informasi yang berguna untuk tujuan tertentu [2]. Pada teks berita tentu menyimpan informasi di dalamnya, dari informasi tersebut kemudian akan diproses untuk menemukan pola data agar dapat dikelompokkan ke dalam kategori-kategori tertentu, proses inilah yang dinamakan dengan klasifikasi [3]. Pada umumnya pemberian label pada proses klasifikasi masih dilakukan dengan caramanual oleh tim ahli pada *dataset* dalam jumlah yang cukup banyak. Oleh karena itu, pada penelitian ini akan membuat sebuah proses klasifikasi otomatis menggunakan metode C4.5 yang akan mengolah hasil pelabelan manual untuk menghasilkan model

terbaik yang nantinya akan digunakan sebagai acuan pada proses pelabelan otomatis berskala besar.

Algoritma C4.5 adalah algoritma yang digunakan untuk membangun metode klasifikasi pohon keputusan[4]. Algoritma C4.5 banyak digunakan dalam proses klasifikasi data dengan atribut numerik dan kategorik. Algoritma C4.5 merupakan pengembangan dari algoritma ID3 yang memiliki kelebihan yaitu kuat terhadap data noise, dapat menangani variabel tipe diskrit dan kontin dapat menangani variabel dengan nilai yang hilang dan dapat memotong cabang pada pohon keputusan [5]. Algoritma C4.5 memiliki input data latih dan data uji. Data latih adalah contoh data yang digunakan dalam proses membangun pohon keputusan yang telah diuji akurasinya, dan data uji adalah bidang data yang digunakan sebagai parameter saat mengklasifikasikan data. Penelitian yang akan dilakukan untuk memperoleh tingkat akurasi yang baik dan pola klasifikasi berita yang menarik, sehingga mendapatkan hasil klasifikasi berita sesuai dengan kategori yang di inginkan[6].

Penelitian ini dilatarbelakangi oleh kesulitan yang dialami oleh Badan Pusat Statistik (BPS) provinsi Riau dalam mencari dan mengelompokkan berita. BPS Provinsi Riau merupakan kementerian yang berperan penting dalam menyediakan data statistik untuk Provinsi Riau. Pengelompokan berita dibagi kedalam 3kategori yaitu demokrasi, ketenagakerjaan, dan kemiskinan. Berita yang dikelompokkan tersebut digunakan sebagai landasan fenomena dari nilai indeks demokrasi, nilai indeks ketenagakerjaan, dan nilai indeks kemiskinan Provinsi Riau. Proses mengkategorikan berita dilakukan secara manual oleh tim ahli di bidang sosial dengan mencari dan membaca setiap berita lalu memberikan label yang menjadi kategori dari setiap berita tersebut. Oleh karena itu, proses pengklasifikasian ini merupakan kegiatan yang sangat memakan waktu dan tenaga sehingga hasil yang didapatkan juga tidak optimal. Dengan alasan tersebut disertai meningkatnya permintaan data dari Badan Pusat Statistik Provinsi Riau, peneliti terdorong untuk melakukan penelitian mengenai penerapan metode klasifikasi otomatis menggunakan algoritma C4.5.

Oleh karena itu, tujuan dalam penelitian ini untuk mengklasifikasikan berita sesuai dengan kategorinya dan mengetahui tingkat akurasi klasifikasi menggunakan metode C4.5, agar dapat membantu tim ahli di bidang sosial Badan Pusat Statistik Provinsi Riau dalam mencari dan pengelompokkan berita.

2. Tinjauan Pustaka

2.1. Berita

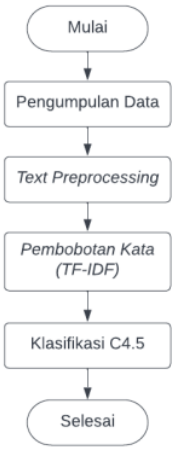
Kata "berita" berasal dari bahasa Sansekerta vrit (ada atau terjadi) atau vritta (peristiwa atau peristiwa). Berita adalah laporan penting, menarik, dan terkini tentang peristiwa, opini, tren, keadaan, dan interpretasi, dan harus segera diketahui publik [7].

2.2. Dasar Teori

Penelitian mengenai topik klasifikasi menggunakan metode C4.5 akan terus berkembang. Hal ini dibuktikan dengan banyaknya penelitian mengenai topik tersebut. Beberapa penelitian mengenai kalsifikasi yang terkait dengan penelitian ini diantaranya seperti penelitian yang dilakukan oleh Muhammad Nur Akbar pada tahun 2021 dengan judul penelitian "Klasifikasi Bibliografi Otomatis Menggunakan C4.5 dan Information Gain" dimana algoritma C4.5 dianggap cukup efisien untuk mengklasifikasikan dokumen, hasil akurasi sebesar 88.58% [8]. Penelitian berikutnya berjudul "Penerapan Data Mining dengan Menggunakan Algoritma C4.5 pada Klasikasi Fasilitas Kesehatan Provinsi di Indonesia" hasil akurasi menunjukkan bahwa algoritma C4.5 dengan cross validation lebih baik dengan nilai 97,5% [9]. Penelitian berikutnya oleh Lukhayu Pritalia pada tahun 2018 dengan judul "Penerapan Algoritma C4.5 untuk Penentuan Ketersediaan Barang E-commerce" menggunakan aplikasi Weka yang berhasil menentukan ketersediaan barang e-commerce dengan presentase keakuratan sebesar 98% [10] . Pada penelitian lain yang berjudul "Penerapan Data Mining Untuk Memprediksi Penerimaan Peserta Didik Baru Menggunakan Algoritma C4.5" oleh Winanjaya dkk pada tahun 2019 mendapatkan hasil bahwa siswa dengan nilai tes tinggi memiliki peluang lebih besar mendapat status diterima dibandingan siswa dengan nilai sedang [11]. Dan beberapa penelitian lainnya [12] [13] [14] yang juga menjadi sumber referensi penelitian ini.

3. MetodePenelitian

Penelitian ini dilakukan di Badan Pusat Statistik Provinsi Riau dengan menggunakan Algoritma C 4.5. Metode penelitian yang dilakukan pada penelitian ini yaitu metode kuantitatif. Pada penelitian kuantitatif menekankan tentang pengujian teori berdasarkan variable yang diukur dan melakukan analisa terhadap data dengan prosedur statistic. Dalam penelitian kuantitatif terdapat batasan dalam lingkup penelitian yang membatasi variabel dan populasi yang digunakan dalam penelitian. Penelitian kuantitatif dilakukan dengan rancangan tahapan yang terstruktur dan sesuai dengan sistematika penelitian ilmiah [15]. Berikut tahapan yang akan dilakukan dalam penelitian ini :



Gambar 1. Tahapan Penelitian

3.1. Pengumpulan Data

Pengumpulan data berita pada penelitian ini dilakukan secara manual terhitung dari tanggal 30 Januari 2022 sampai dengan 10 Maret 2022 yang dikumpulkan dari beberapa portalberita terpercaya yang memberitakan seputar Provinsi Riau. Data yang dikumpulkan berjumlah 510 data berita, selanjutnya dilakukan proses pelabelan secara manual oleh para ahli dari bidang sosial di Badan Pusat Statistik Provinsi Riau. Dari 510 berita tersebut, sebanyak 170 berita berkategori demokrasi, 170 berita berkategori kemiskinan, dan 170 berita berkategori ketenagakerjaan. Dokumen berita disimpan dalam bentuk format *google spreadsheet*. Berikut ini merupakan contoh *dataset*berita yang telah dikumpulkan dan dilabeli:

Berita	Kelas
Dewan Perwakilan Rakyat Daerah (DPRD) Kabupaten Pelalawan segera mengagendakan Rapat Paripurrna dengan agenda pelantikan Pergantian Antar Waktu (PAW) Wakil Ketua DPRD Anton Sugianto, S.Ud digantikan oleh Faizal, SE M.Si.	Demokrasi
Pemerintah Kota (Pemko) Pekanbaru sedang mengumpulkan data masyarakat miskin positif Covid-19 yang sedang isolasi di fasilitas pemerintah atau sedang diopname. Namun, kendalanya data itu sampai hari ini belum disampaikan oleh pihak kelurahan kepada Dinas Sosial (Dinsos) Kota Pekanbaru.	Kemiskinan
Kepala Dinas Sosial Tenaga Kerja dan Transmigrasi (Disosnakertrans) Siak Nurmansyah melayangkan surat himbauan kepada Badan Operasi Bersama (BOB) PT Bumi Siak Pusako (BSP)-Pertamina Hulu, agar menunda pengurangan tenaga kerja.	Ketenagakerjaan

Tabel 1. Dataset Berita

Dataset yang berjumlah 510 data akan dibagi menjadi 2 bagian yaitu data latih dan data uji dengan menggunakan 3 skenario perbandingan data yaitu 70% data latih dan 30% data uji,80% data latih dan 20% data uji, lalu 90% data latih dan 10% data uji. Berikut rincian perbandingan dataset yang digunakan:

Data Berita	Kelas	Pembagian Data Latih dan Data uji					
		Training	Testing	Training	Testing	Training	Testing
		70%	30%	80%	20%	90%	10%
510	Demokrasi	119	51	136	34	153	17
	Kemiskinan	119	51	136	34	153	17
	Ketenagakerjaan	119	51	136	34	153	17
Jumlah		357	153	408	102	459	51

Tabel 2. Skenario Pembagian Data

3.2. *Text Preprocessing*

Setelah proses pengumpulan data selesai dilakukan, langkah selanjutnya adalah Text Preprocessing. Text *Preprocessing* merupakan proses yang sangat penting dalam klasifikasi. Tujuan dari preprocessing teks adalah untuk membersihkan data dari komponen yang tidak dibutuhkan, menyeragamkan bentuk kata dan mengurangi jumlah kata agar mempermudah proses klasifikasi. Tahapan pada Text Preprocessing diantaranya:

1. *Cleaning* (Tahapan *cleaning* dilakukan pembersihan kata yaitu menghilangkan karakter, simbol, atau identitas pengguna yang tidak diperlukan seperti (!@#$%^&*():{}.,?~/[]), *URL,* angka, dan *emoticon*.)
2. *Case Folding* (Tahapan *case folding* dilakukan pengubahan seluruh huruf pada data menjadi huruf kecil *(lowercase)* untuk menyeragamkan kata pada berita).
3. *Tokenizing* (Tahapan *tokenizing* dilakukan pada data berita yang awalnya merupakan sebuah kalimat dipecah menjadi potongan-potongan kata atau disebut dengan token).
4. Normalisasi (Tahapan *normalisasi* dilakukan perubahan kata yang tidak baku atau yang salah ejaannya menjadi kata baku sesuai KBBI dengan menggunakan kamus normalisasi yang dibuat manual berdasarkan pengecekkan data secara manual).
5. *Removal Stopword* (Tahapan *removal stopword* dilakukan untuk menghilangkan kata-kata yang tidak berpengaruh dalam proses klasifikasi contohnya seperti kata hubung. Proses penghapusan kata-kata tersebut berdasarkan filtering menggunakan *library* NLTK untuk bahasa Indonesia dan kamus *stopword* yang dibuat sendiri secara manual).
6. *Stemming* (Tahapan *stemming* dilakukan untuk mendapatkan kata dasar dengan menghilangkan imbuhan awalan, akhiran, sisipan, dan *confixes* (kombinasi awalan dan akhiran) sesuai panduan KBBI. Pada proses *stemming* ini menggunakan kelas *StemmerFactory* dari *library* Sastrawi).

3.3. Pembobotan Kata (TF-IDF)

Setelah langkah preprocessing selesai selanjutnya yaitu tahap pembobotan kata. Dalam penelitian ini, bobot diberikan menggunakan TF-IDF (Term Frequency – Inverse Document Frequency). Metode TF-IDF dipilih pada tahap ini karena memiliki nilai presisi dan recall yang lebih baik, serta waktu yang dibutuhkan untuk eksekusi lebih cepat dibandingkan metode lainnya. Berikut adalah langkah-langkah dalam proses pembobotan kata menggunakan metode TF-IDF:

1. Persiapkan data yang sudah diolah pada tahapan *preprocessing*.
2. Menghitung jumlah kata yang muncul dalam sebuah dokumen *(TF)*.

$$tf_{ij} = \frac{f_d}{max\, f_d(j)}$$

3. Menghitung jumlah dokumen dimana suatu kata itu muncul *(IDF)*.

$$idf_{ij} = \frac{D}{df_i}$$

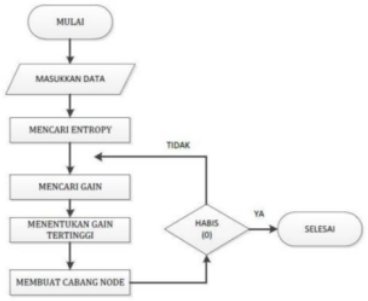
4. Menghitung *TF-IDF*.

$$w_{ij} = tf_{ij}\ x\ idf_{ij}$$

5. Mendapatkan hasil data.

3.3. Klasifikasi C4.5

Setelah tahap preprocessing selanjutnya adalah proses klasifikasi. Pada penelitian ini digunakan Algoritma C4.5 karena algoritma ini dikenal sederhana dan mudah dipahami karena berupa pohon keputusan serta memiliki performa yang baik dalam mengklasifikasi dokumen. Berikut langkah-langkah dalam pembuatan pohon keputusan menggunakan algoritma C4.5 :



Gambar 2. Tahapan membuat pohon keputusan C4.5

Proses dari pohon keputusan adalah mengubah bentuk data (tabel) menjadi model pohon, mengubah model pohon menjadi aturan, dan menyederhanakan aturan [16]. Berikut adalah langkah-langkah untuk membuat pohon keputusan menggunakan algoritma C4.5:

1. Menentukan data training.
2. Menghitung nilai entropy pada seluruh data atau entropy total.

$$Entropy = - \sum pi \times log2pi \quad k\ i=1$$

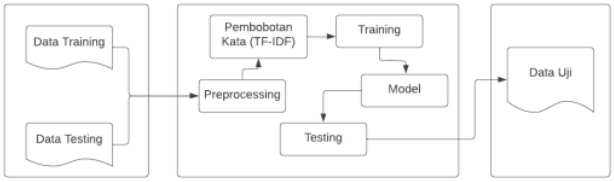
3. Menghitung nilai gain pada masing-masing atribut.

$$Gain(S,A) = Entropy\ (S) - \sum |Si\ | |S\ | \times Entropy(Si\ ) \quad k\ i=1$$

4. Menghitung nilai gain ratio pada masing-masing atribut.

$$GainRatio(S,A) = \frac{Gain(S,A)}{Split(S,A)}$$

5. Memilih atribut yang akan digunakan sebagai node berdasarkan nilai gain ratio tertinggi.
6. Mengecek apakah semua atribut sudah terbentuk pada pohon, jika belum maka ulangi proses 2-5, jika sudah maka lanjut pada proses selanjutnya.
7. Aturan keputusan di-generate mengikuti pohon yang telah dibentuk sebelumnya.



Gambar 3. Tahapan Klasifikasi

Pada gambar 4. diatas terdapat tiga proses yang dilakukan, yaitu:

1. Input
   Data yang akan diinputkan adalah keseluruhan data berita yang telah dikumpulkan sebanyak 510 data dan telah dilakukan proses pelabelan secara manual. Selanjutnya data berita akan dibagi menjadi data latih dan data uji sesuai skenario 70% : 30%, 80% : 20%, dan 90% : 10% untuk mendapatkan nilai akurasi terbaik.
2. Proses Klasifikasi C4.5
   Tahapan mulai dari mempersiapkan dataset dengan melakukan *text preprocessing*, pembobotan kata (TF-IDF) sampai ke pembentukan pemodelan C4.5.
3. Output
   Data yang akan dihasilkan berupa data berita yang telah diprediksi oleh pemodelan C4.5 dengan kelas Demokrasi, Kemiskinan dan Ketenagakerjaan.

4. Hasil dan Pembahasan

Berikut merupakan hasil dari setiap proses dalam penelitian ini :

4.1. *Text Preprocessing*

1. *Cleaning*
   Hasil dari tahapan cleaning pada Tabel 3 dibawah ini:

Hasil Cleaning
Dewan Perwakilan Rakyat Daerah DPRD Kabupaten Pelalawan segera mengagendakan Rapat Paripurna dengan agenda pelantikan Pergantian Antar Waktu PAW Wakil Ketua DPRD Anton Sugianto S Ud digantikan oleh Faizal SE M Si
Pemerintah Kota Pemko Pekanbaru sedang mengumpulkan data masyarakat miskin positif Covid yang sedang isolasi di fasilitas pemerintah atau sedang diopname Namun kendalanya data itu sampai hari ini belum disampaikan oleh pihak kelurahan kepada Dinas Sosial Dinsos Kota Pekanbaru

Kepala Dinas Sosial Tenaga Kerja dan Transmigrasi Disosnakertrans Siak Nurmansyah melayangkan surat himbauan kepada Badan Operasi Bersama BOB PT Bumi Siak Pusako BSP Pertamina Hulu agar menunda pengurangan tenaga kerja

Tabel 3. Hasil Cleaning

2. *Case Folding*

Hasil dari tahapan cleaning pada Tabel 4 dibawah ini:

Hasil Case Folding
dewan perwakilan rakyat daerah dprd kabupaten pelalawan segera mengagendakan rapat paripurrna dengan agenda pelantikan pergantian antar waktu paw wakil ketua dprd anton sugianto s ud digantikan oleh faizal se m si
pemerintah kota pemko pekanbaru sedang mengumpulkan data masyarakat miskin positif covid yang sedang isolasi di fasilitas pemerintah atau sedang diopname namun kendalanya data itu sampai hari ini belum disampaikan oleh pihak kelurahan kepada dinas sosial dinsos kota pekanbaru
kepala dinas sosial tenaga kerja dan transmigrasi disosnakertrans siak nurmansyah melayangkan surat himbauan kepada badan operasi bersama bob pt bumi siak pusako bsp pertamina hulu agar menunda pengurangan tenaga kerja

Tabel 4. Hasil Case Folding

3. *Tokenizing*

Hasil dari tahapan cleaning pada Tabel 5 dibawah ini:

D1	**D2**	**D3**
'dewan',	'pemerintah'	'kepala'
'perwakilan'	'kota'	'dinas'
'rakyat'	'pemko'	'sosial'
'daerah'	'pekanbaru'	'tenaga'
'dprd'	'sedang'	'kerja'
'kabupaten'	'mengumpulkan'	'dan'
'pelalawan'	'data'	'transmigrasi'
'segera'	'masyarakat'	'disosnakertrans'
'mengagendakan'	'miskin'	'siak'
'rapat'	'positif'	'nurmansyah'
'paripurrna'	'covid'	'melayangkan'
'dengan'	'yang'	'surat'
'agenda'	'sedang'	'himbauan'
'pelantikan'	'isolasi'	'kepada'
'pergantian'	'di'	'badan'
'antar'	'fasilitas'	'operasi'
'waktu'	'pemerintah'	'bersama'
'paw'	'atau'	'bob'
'wakil'	'sedang'	'pt'
'ketua'	'diopname'	'bumi'
'dprd'	'namun'	'siak'
'anton'	'kendalanya'	'pusako'
'sugianto'	'data'	'bsp'
's'	'itu'	'pertamina'
'ud'	'sampai'	'hulu'
'digantikan'	'hari'	'agar'
'oleh'	'ini'	'menunda'
'faizal'	'belum'	'pengurangan'
'se'	'disampaikan'	'tenaga'
'm'	'oleh'	'kerja'
'si'	'pihak'	
	'kelurahan'	
	'kepada'	
	'dinas'	
	'sosial'	
	'dinsos'	
	'kota'	
	'pekanbaru'	

Tabel 5. Hasil Tokenizing

4. Normalisasi

Hasil dari tahapan cleaning pada Tabel 6 dibawah ini:

D1	D2	D3
'dewan',	'pemerintah'	'kepala'
'perwakilan'	'kota'	'dinas'
'rakyat'	'pemko'	'sosial'
'daerah'	'pekanbaru'	'tenaga'
'dprd'	'sedang'	'kerja'
'kabupaten'	'mengumpulkan'	'dan'
'pelalawan'	'data'	'transmigrasi'
'segera'	'masyarakat'	'disosnakertrans'
'mengagendakan'	'miskin'	'siak'
'rapat'	'positif'	'nurmansyah'
'paripurna'	'covid'	'melayangkan'
'dengan'	'yang'	'surat'
'agenda'	'sedang'	'himbauan'
'pelantikan'	'isolasi'	'kepada'
'pergantian'	'di'	'badan'
'antar'	'fasilitas'	'operasi'
'waktu'	'pemerintah'	'bersama'
'paw'	'atau'	'bob'
'wakil'	'sedang'	'pt'
'ketua'	'diopname'	'bumi'
'dprd'	'namun'	'siak'
'anton'	'kendalanya'	'pusako'
'sugianto'	'data'	'bsp'
's'	'itu'	'pertamina'
'ud'	'sampai'	'hulu'
'digantikan'	'hari'	'agar'
'oleh'	'ini'	'menunda'
'faizal'	'belum'	'pengurangan'
'se'	'disampaikan'	'tenaga'
'm'	'oleh'	'kerja'
'si'	'pihak'	
	'kelurahan'	
	'kepada'	
	'dinas'	
	'sosial'	
	'dinsos'	
	'kota'	
	'pekanbaru'	

Tabel 6. Hasil Normalisasi

5. *Stopword Removal*

Hasil dari tahapan cleaning pada Tabel 7 dibawah ini:

D1	D2	D3
'dewan',	'pemerintah'	'kepala'
'perwakilan'	'kota'	'dinas'
'rakyat'	'pemko'	'sosial'
'daerah'	'pekanbaru'	'tenaga'
'dprd'	'mengumpulkan'	'kerja'
'kabupaten'	'data'	'transmigrasi'
'pelalawan'	'masyarakat'	'disosnakertrans'
'mengagendakan'	'miskin'	'siak'
'rapat'	'positif'	'nurmansyah'
'paripurna'	'covid'	'melayangkan'
'agenda'	'isolasi'	'surat'
'pelantikan'	'fasilitas'	'himbauan'
'pergantian'	'pemerintah'	'badan'
'wakil'	'diopname'	'operasi'
'ketua'	'kendalanya'	'bersama'

'dprd'	'data'	'bumi'
'anton'	'kelurahan'	'siak'
'sugianto'	'dinas'	'pusako'
'digantikan'	'sosial'	'bsp'
'faizal'	'dinsos'	'pertamina'
	'kota'	'hulu'
	'pekanbaru'	'menunda'
		'pengurangan'
		'tenaga'
		'kerja'

Tabel 7. Hasil Stopword Removal

6. *Stemming*

Hasil dari tahapan cleaning pada Tabel 8 dibawah ini:

D1	**D2**	**D3**
'dewan',	'pemerintah'	'kepala'
'wakil'	'kota'	'dinas'
'rakyat'	'pemko'	'sosial'
'daerah'	'pekanbaru'	'tenaga'
'dprd'	'kumpul'	'kerja'
'kabupaten'	'data'	'transmigrasi'
'pelalawan'	'masyarakat'	'disosnakertrans'
'agenda'	'miskin'	'siak'
'rapat'	'positif'	'nurmansyah'
'paripurrna'	'covid'	'layang'
'agenda'	'isolasi'	'surat'
'lantik'	'fasilitas'	'himbau'
'ganti'	'pemerintah'	'badan'
'wakil'	'diopname'	'operasi'
'ketua'	'kendala'	'bersama'
'dprd'	'data'	'bumi'
'anton'	'lurah'	'siak'
'sugianto'	'dinas'	'pusako'
'ganti'	'sosial'	'bsp'
'faizal'	'dinsos'	'pertamina'
	'kota'	'hulu'
	'pekanbaru'	'tunda'
		'kurang'
		'tenaga'
		'kerja'

Tabel 8. Hasil Stemming

4.2. Pembobotan Kata (TF-IDF)

Setelah tahapan preprocessing selesai selanjutnya pembobotan kata menggunakan TF-IDF. Berikut Hasil perhitungan TF-IDF pada Tabel 9 dibawah ini:

Term	**TF**			**DF**	**IDF**	**TF-IDF**		
	D1	**D2**	**D3**			**D1**	**D2**	**D3**
Agenda	2	0	0	1	*ln*(4/2)+1 = 1.693	3.386	0	0
Anton	1	0	0	1	*ln*(4/2)+1 = 1.693	1.693	0	0
Badan	0	0	1	1	*ln*(4/2)+1 = 1.693	0	0	1.693
Bsp	0	0	1	1	*ln*(4/2)+1 = 1.693	0	0	1.693
Bumi	0	0	1	1	*ln*(4/2)+1 = 1.693	0	0	1.693
Covid	0	1	0	1	*ln(4/2)+1 = 1.693*	0	1.693	0
Daerah	1	0	0	1	*ln(4/2)+1 = 1.693*	1.693	0	0
Data	0	2	0	1	*ln(4/2)+1 = 1.693*	0	3.386	0
Dewan	1	0	0	1	*ln(4/2)+1 = 1.693*	1.693	0	0
Dinas	0	1	1	2	*ln(4/2)+1 = 1.287*	0	1.287	1.287
Dinsos	0	1	0	1	*ln(4/2)+1 = 1.693*	0	1.693	0
disosnakertrans	0	0	1	1	*ln(4/2)+1 = 1.693*	0	0	1.693
Dprd	2	0	0	1	*ln(4/2)+1 = 1.693*	3.386	0	0
Faizal	1	0	0	1	*ln(4/2)+1 = 1.693*	1.693	0	0

Fasilitas	0	1	0	1	*ln(4/2)+1 = 1.693*	0	1.693	0
ganti	2	0	0	1	*ln(4/2)+1 = 1.693*	3.386	0	0
Himbauan	0	0	1	1	*ln(4/2)+1 = 1.693*	0	0	1.693
Hulu	0	0	1	1	*ln(4/2)+1 = 1.693*	0	0	1.693
Isolasi	0	1	0	1	*ln(4/2)+1 = 1.693*	0	1.693	0
Kabupaten	1	0	0	1	*ln(4/2)+1 = 1.693*	1.693	0	0
Kendala	0	1	0	1	*ln(4/2)+1 = 1.693*	0	1.693	0
Kepala	0	0	1	1	*ln(4/2)+1 = 1.693*	0	0	1.693
Kerja	0	0	2	1	*ln(4/2)+1 = 1.693*	0	0	3.386
Ketua	1	0	0	1	*ln(4/2)+1 = 1.693*	1.693	0	0
Kota	0	2	0	1	*ln(4/2)+1 = 1.693*	0	3.386	0
Kumpul	0	1	0	1	*ln(4/2)+1 = 1.693*	0	1.693	0
Kurang	0	0	1	1	*ln(4/2)+1 = 1.693*	0	0	1.693
Lantik	1	0	0	1	*ln(4/2)+1 = 1.693*	1.693	0	0
Laying	0	0	1	1	*ln(4/2)+1 = 1.693*	0	0	1.693
Lurah	0	1	0	1	*ln(4/2)+1 = 1.693*	0	1.693	0
Masyarakat	0	1	0	1	*ln(4/2)+1 = 1.693*	0	1.693	0
Miskin	0	1	0	1	*ln(4/2)+1 = 1.693*	0	1.693	0
Nurmansyah	0	0	1	1	*ln(4/2)+1 = 1.693*	0	0	1.693
Operasi	0	0	1	1	*ln(4/2)+1 = 1.693*	0	0	1.693
Opname	0	1	0	1	*ln(4/2)+1 = 1.693*	0	1.693	0
Paripurrna	1	0	0	1	*ln(4/2)+1 = 1.693*	1.693	0	0
Pekanbaru	0	2	0	1	*ln(4/2)+1 = 1.693*	0	3.386	0
Pelalawan	1	0	0	1	*ln(4/2)+1 = 1.693*	1.693	0	0
pemko	0	1	0	1	*ln(4/2)+1 = 1.693*	0	1.693	0
Perintah	0	2	0	1	*ln(4/2)+1 = 1.693*	0	3.386	0
Pertamina	0	0	1	1	*ln(4/2)+1 = 1.693*	0	0	1.693
Positif	0	1	0	1	*ln(4/2)+1 = 1.693*	0	1.693	0
Pusako	0	0	1	1	*ln(4/2)+1 = 1.693*	0	0	1.693
Rakyat	1	0	0	1	*ln(4/2)+1 = 1.693*	1.693	0	0
Rapat	1	0	0	1	*ln(4/2)+1 = 1.693*	1.693	0	0
Siak	0	0	2	1	*ln(4/2)+1 = 1.693*	0	0	3.386
Social	0	1	1	2	*ln(4/2)+1 = 1.287*	0	1.287	1.287
Sugianto	1	0	0	1	*ln(4/2)+1 = 1.693*	1.693	0	0
Surat	0	0	1	1	*ln(4/2)+1 = 1.693*	0	0	1.693
Tenaga	0	0	2	1	*ln(4/2)+1 = 1.693*	0	0	3.386
Transmigrasi	0	0	1	1	*ln(4/2)+1 = 1.693*	0	0	1.693
Tunda	0	0	1	1	*ln(4/2)+1 = 1.693*	0	0	1.693
Wakil	2	0	0	1	*ln(4/2)+1 = 1.693*	3.386	0	0

Tabel 9. Hasil Perhitungan TF-IDF

4.3. Klasifikasi C4.5

Data berita yang telah melalui tahapan text preprocessing dan pembobotan kata (tf-idf) selanjutnya akan melalui proses klasifikasi menggunakan metode C4.5 untuk menghasilkan model yang dapat mengklasifikasikan data berita baru tanpa pelabelan manual. Pada tahap ini proses pembelajaran dan pelatihan dilakukan. Proses pembelajaran dengan melatih model menggunakan dataset yang sudah diberi label sedangkan pada proses pelatihan dilakukan menggunakan data yang belu diberi label. Hasil dari klasifikasi C4.5 pada penelitian ini sudah cukup baik dengan hasil akurasi yang cukup tinggi, algoritma ini dianggap cocok dalam mengklasifikasikan data berita.

4.4. Pengujian Confusion Matrix

1. Pengujian terhadap 90% data latih dan 10% data uji

Berikut adalah hasil pengujian confusion matrix menggunakan 90% data latih dan 10% data uji pada Tabel 10 dibawah ini:

Predicted Classification	*Actual Classification*		
	Demokrasi Positive	Kemiskinan Positive	KetenagakerjaanPositive
DemokrasiNegative	10	2	1
KemiskinanNegative	1	20	1
KetenagakerjaanNegative	3	0	13

Tabel 10. Hasil Confusion Matrix pada data 90%:10%

Perhitungan akurasi $= \frac{10+20+13}{10+2+1+1+20+1+3+0+13} \; x \; 100\%$
$= \frac{43}{51} \; x \; 100\%$
$= 84\%$

2. Pengujian terhadap 80% data latih dan 20% data uji
Berikut adalah hasil pengujian confusion matrix menggunakan 80% data latih dan 20% data uji pada Tabel 11 dibawah ini:

Predicted Classification	*Actual Classification*		
	Demokrasi Positive	Kemiskinan Positive	KetenagakerjaanPositive
DemokrasiNegative	20	0	7
KemiskinanNegative	1	30	1
KetenagakerjaanNegative	9	1	33

Tabel 11. Hasil Confusion Matrix pada data 80%:20%

Perhitungan akurasi $= \frac{20+30+33}{20+0+7+1+30+1+9+1+33} \; x \; 100\%$
$= \frac{83}{102} \; x \; 100\%$
$= 81\%$

3. Pengujian terhadap 70% data latih dan 30% data uji
Berikut adalah hasil pengujian confusion matrix menggunakan 70% data latih dan 30% data uji pada Tabel 12 dibawah ini:

Predicted Classification	*Actual Classification*		
	Demokrasi Positive	Kemiskinan Positive	KetenagakerjaanPositive
DemokrasiNegative	31	2	5
KemiskinanNegative	4	48	3
KetenagakerjaanNegative	12	2	46

Tabel 12. Hasil Confusion Matrix pada data 70%:30%

Perhitungan akurasi $= \frac{31+48+46}{31+2+5+4+48+3+12+2+46} \; x \; 100\%$
$= \frac{125}{153} \; x \; 100\%$
$= 81\%.$

5. Kesimpulan

Berdasarkan hasil dari penelitian yang telah dilakukan, dapat disimpulkan beberapa hal sebagaiberikut:
1. Metode *C4.5* terbukti dapat digunakan dalam proses klasifikasi berita.
2. Hasil akurasi tertinggi yang didapatkan dari proses klasifikasi yaitu 84% menggunakan pengujian *confusion matrix* dengan pembagian data latih 90% dan data uji 10% dari *dataset* yang digunakan.

Daftar Pustaka

[1] E. Y. Hidayat and M. A. Rizqi, "Klasifikasi Dokumen Berita Menggunakan Algoritma Enhanced Confix Stripping Stemmer dan Naïve Bayes Classifier," *J. Nas. Teknol. dan Sist. Inf.*, vol. 6, no. 2, pp. 90–99, 2020.
[2] N. M. G. D. Purnamasari, M. A. Fauzi, Indriarti, and L. S. Dewi, "Identifikasi Tweet Cyberbullying pada Aplikasi Twitter menggunakan Metode Support Vector Machine (SVM) dan Information Gain (IG) sebagai Seleksi Fitur," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 2, no. 11, pp. 5326–5332, 2018.
[3] V. Chandani and R. S. Wahono, "Komparasi Algoritma Klasifikasi Machine Learning Dan Feature Selection pada Analisis Sentimen Review Film," *J. Intell. Syst.*, vol. 1, no. 1, pp. 55–59, 2015.
[4] H. Dhika and F. Destiawati, "Penerapan Algoritma C45 Untuk Penilaian Karyawan Pada Restoran Cepat Saji," no. September, pp. 55–59, 2018.
[5] Y. Mardi, "Data Mining : Klasifikasi Menggunakan Algoritma C4.5," *Edik Inform.*, vol. 2, no. 2, pp. 213–219, 2017.
[6] F. F. Harryanto and S. Hansun, "Penerapan Algoritma C4.5 untuk Memprediksi Penerimaan Calon Pegawai Baru di PT WISE," *Maret*, vol. 3, no. 2, p. 95, 2017.

[7] L. Septia and D. Br, *BUKU JURNALISTIK PDF-min*, no. October. 2020.
[8] M. N. AKBAR, "Klasifikasi Bibliografi Otomatis Menggunakan C4.5 Dan Information Gain," *J. INSTEK (Informatika Sains dan Teknol.*, vol. 6, no. 1, p. 46, 2021.
[9] C. Algoritma, A. Carolina, K. Ade, and K. Kunci, "Penerapan Data Mining Dengan Menggunakan Algoritma C4.5 Pada Klasifikasi Fasilitas Kesehatan Provinsi Di Indonesia," *J. Ilm. Komputasi*, vol. 19, no. 1, pp. 27–38, 2020.
[10] G. Lukhayu Pritalia, "Penerapan Algoritma C4.5 untuk Penentuan Ketersediaan Barang E-commerce," *Indones. J. Inf. Syst.*, vol. 1, no. 1, pp. 47–56, 2018.
[11] R. Winanjaya, F. Amir, and R. Doni, "Penerapan Data Mining Untuk Memprediksi Penerimaan Peserta Didik Baru Menggunakan Algoritma C4.5," *Pros. Semin. Nas. Ris. Inf. Sci.*, vol. 1, no. September, p. 1, 2019.
[12] M. Firmansyah and R. Aufany, "Implementasi Metode Decision Tree Dan Algoritma C4.5 Untuk Klasifikasi Data Nasabah Bank," *Infokam*, vol. XII, no. 1, pp. 1–12, 2016.
[13] S. Haryati, A. Sudarsono, and E. Suryana, "Implementasi Data Mining Untuk Memprediksi Masa Studi Mahasiswa Menggunakan Algoritma C4.5 (Studi Kasus: Universitas Dehasen Bengkulu)," *J. Media Infotama*, vol. 11, no. 2, pp. 130–138, 2015.
[14] U. N. Tantyoko. Adiwijaya. & Wisesty, "35-Article Text-89-1-10-20190908.pdf." pp. 97–113, 2019.
[15] Hardani *et al.*, *Metode Penelitian Kualitatif & Kuantitatif*, no. April. 2020.
[16] R. Sari, "Komparasi Algoritma Support Vector Machine, Naïve Bayes Dan C4.5 untuk Klasifikasi SMS," *IJCIT(Indonesia J. Comput. Infomation Technol.*, vol. 2, no. 2, pp. 7–13, 2017.

Jurnal_Nas_Terakreditasi_q.pdf

ORIGINALITY REPORT

8%	8%	6%	5%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Higher Education Commission Pakistan Student Paper	5%
2	www.semanticscholar.org Internet Source	2%
3	journal.uin-alauddin.ac.id Internet Source	2%

Exclude quotes	On	Exclude matches	< 2%
Exclude bibliography	On		