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

Klasifikasi Berita Menggunakan Metode *K-Nearest Neighbor*

Rahmah Miya Juwita[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: 11850120438@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 - Meningkatnya minat masyarakat dalam mengakses berita, khususnya berita online, menuntut redaktur dan situs portal berita untuk memberikan liputan dan berita yang berkualitas. Selain itu, klasifikas berita yang ada masih tergolong umum dapat menjadi kendala yang dialami pembaca. jika pembaca ingin melihat kategori berita yang lebih spesifik, mereka harus menyaring berita tersebut secara manual. Hal ini juga terjadi di bidang sosial Badan Pusat Statistik Provinsi Riau yang kesulitan mencari berita tentang Provinsi Riau. Oleh karena itu, proses klasifikasi berita menggunakan metode k-nearest neighbor menjadi hal yang krusial untuk dilakukan. Jumlah berita yang digunakan dalam penelitian ini berjumlah 510 data dengan tiga kategori yaitu demokrasi, kemiskinan, dan ketenagakerjaan. Proses klasifikasi berita dalam penelitian ini meliputi: pengumpulan data, pelabelan manual, preprocessing teks, pembobotan kata, dan klasifikasi memakai metode k-nearest neighbor. Selain itu, cosinus similarity juga digunakan untuk meningkatkan nilai akurasi. Nilai akurasi tertinggi yang diperoleh pada penelitian ini adalah 87% menggunakan nilai k = 3 dengan distribusi data uji 20% & dan data latih dari 80%. Dari penelitian ini dapat diambil kesimpulan bahwa metode K-Nearest Neighbor dapat bekerja dengan baik dalam proses klasifikasi berita.
Kata kunci: *Badan Pusat Statistik, Berita, Cosine Similarity, Klasifikasi, K-Nearest Neighbor*

Abstract **-** The increasing of public interest in accessing news, especially online news, requires editors and news portal sites to provide quality coverage and news. In addition, the grouping of news that still classified as a general can be an obstacle experienced by readers. if the reader wants to see a more specific category of news, they must filter the news manually. This is also happened in the social sector of Badan Pusat Statistik Provinsi Riau, which has trouble when finding news about Riau Province. Therefore, the news classification process using the k-nearest neighbor method is a crucial thing to do. The number of news stories used in this study amounted to 510 data with three categories, democracy, poverty, and employment. The news classification process in this study includes: data collection, manual labeling, text preprocessing, word weighting, and classification using k-nearest neighbor method. Besides that, cosine similarity is also used to increase the accuracy value. The highest accuracy values obtained in this study were 87% using a values of k = 3 with distribution of test data of 20% & and training data of 80%. From this research, it can be concluded that the K-Nearest Neighbor method works well in the news classification process.
Keywords: *Badan Pusat Statistik, Cosine Similarity, Classification, K-Nearest Neighbor, News*

1. Pendahuluan

Berita adalah sebuah informasi mengenai hal-hal yang sedang terjadi, disampaikan secara lisan, di internet, di media cetak, ataupun di radio kepada seseorang atau sejumlah besar orang. Dimuat pada Kamus Besar Bahasa Indonesia yang dimaksud dengan berita ialah cerita mengenai suatu kejadian yang hangat yang bersifat pemberitahuan atau pengumuman. Dikarenakan perkembangan teknologi yang semakin maju saat ini, membaca berita tidak hanya melalui media cetak saja melainkan dapat di akses secara daring atau digital. Berdasarkan laporan Reuters Institute yang di lansir dalam situs databooks ,kebanyakan masyarakat Indonesia menggunakan berbagai macam media untuk mengumpulkan informasi atau berita. Sekitar 89% responden yang telah disurvei, menggunakan media online untuk mendapatkan informasi terbaru salah satunya adalah situs berita daring [1]. Berita ini dapat diakses melalui beberapa situs terkenal seperti Kompas.com, Detik.com dan lain sebagainya.

Berita yang dimuat pada media elektronik atau situs berita daring biasanya dikelompokkan berdasarkan berbagai kategori seperti ekonomi, olahraga, dan lain sebagainya. Meskipun beberapa situs telah membuat kategori berita secara umum namun, ketika pembaca ingin mencari informasi didalam kategori tersebut secara spesifik, berita harus disaring secara manual dari kategori yang di pilih kemudian diklasifikasikan menjadi sub kategori yang lebih detail. Hal ini menjadi hambatan bagi pembaca dikarenakan jumlah berita yang terus meningkat. Oleh karenanya diperlukan proses klasifikasi berita menggunakan text mining.

Text mining ialah cabang dari data mining yang digunakan untuk menganalisa data yang berbentuk dokumen teks. Sebelum dilakukan proses analisa terhadap data teks memakai metode pada text mining perlu dilakukan praproses teks antara lain merupakan tokenizing, case folding, stopwords, normalisasi, dan stemming. Sesudah dilakukan praproses teks hal yang dilakukan berikutnya adalah pemberian bobot dan

klasifikasi pada setiap kategori. Klasifikasi adalah suatu metode yang dipergunakan dalam memprediksi kategori atau kelas dari suatu data yang sudah didefinisikan sebelumnya [2]. Penelitian yang akan dilakukan untuk proses klasifikasi berita dengan memakai metode K-Nearest Neighbor. Untuk memperoleh tingkat akurasi yang baik dan pola klasifikasi berita yang menarik, sehingga mendapatkan hasil klasifikasi berita sesuai dengan kategori yang diinginkan digunakan juga cosine similarity sebagai penunjang metode k-nearest neighbor.

Penelitian ini didasari oleh hambatan yang dialami oleh bidang sosial di Badan Pusat Statistik Provinsi Riau dalam mencari dan mengelompokkan berita yang terjadi di Provinsi Riau menjadi 3 kategori yakni demokrasi, ketenagakerjaan, dan kemiskinan. Proses pencarian dan pengelompokkan berita ini juga masih menggunakan metode scrapping manual dengan mencari berita satu persatu pada portal berita dan koran, kemudian diberikan label secara manual pada berita dengan jumlah yang cukup banyak. Sehingga proses klasifikasi berita tidak mendapatkan hasil yang optimal, karena sebagian besar proses yang dilakukan masih secara manual dengan harus membaca isi berita tersebut satu persatu, yang membutuhkan waktu serta tenaga yang cukup besar. Berita yang diklasifikasikan tersebut digunakan sebagai landasan fenomena dari nilai indeks demokrasi, nilai indeks ketenagakerjaan, dan nilai indeks kemiskinan di Provinsi Riau.

Oleh karenanya, tujuan dilakukannya penelitian ini ialah untuk melakukan klasifikasi berita sesuai dengan kategori yang sudah disebutkan sebelumnya serta untuk memperoleh nilai akurasi dari proses klasifikasi metode k-nearest neighbor, agar dapat membantu bidang sosial di Badan Pusat Statistik Provinsi Riau dalam mencari dan pengelompokkan berita unutk dijadikan landasan fenomena yang terjadi di lapangan di daerah Provinsi Riau berdasarkan dari nilai indeks demokrasi, nilai indeks ketenagakerjaan, dan nilai indeks kemiskinan Provinsi Riau.

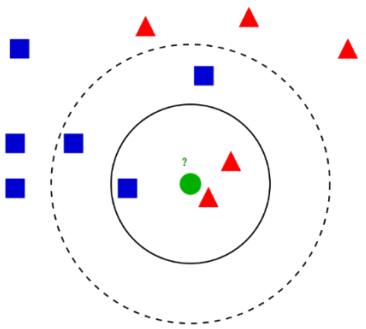
2. Tinjauan Pustaka.

2.1. *Teks Mining*

Teks mining ialah suatu ilmu mengenai proses mengolah teks dokumen pada jumlah besar yang terdapat berdasarkan waktu ke waktu dengan memakai beberapa analisis, tujuan pengolahan teks ialah untuk mengetahui dan mengekstrak informasi yang bermanfaat pada data dengan pola menarik. Dalam kasus text mining, asal data yang dipakai merupakan deretan atau koleksi dokumen yang tidak terstuktur dan memerlukan pengelompokan untuk diketahui informasi yang sejenis [3]

2.2. *K-Nearest Neighbor*

K-Nearest Neighbor (K-NN) ialah sebuah metode yang melakukan proses klasifikasi terhadap suatu data dengan cara pembelajaran jarak paling dekat antara data satu dan lainnya[4].



Gambar 1. Representasi Algoritma K-Nearest Neighbor

Pada gambar 1 diatas data yang sudah ada atau data yang dijadikan sebagai data latih di presentasikan dalam bentuk persegi dan segitiga, sedangkan data uji nya di presentasikan dalam bentuk lingkaran. Dari gambar dapat kita simpulkan bahwa data uji tersebut termasuk ke dalam data latih yang di presentasikan dalam bentuk segtiga. Hal ini didasarkan oleh cara kerja K-NN yaitu klasifikasi sesuai tetangga terdekat. Metode klasifikasi *K-Nearest Neighbor* ini memiliki kelebihan dan kekurangan. Kelebihan dari metode ini yakni tangguh jika dipergunakan untuk *training data* yang mempunyai *noise* yang banyak serta berskala yang besar. Adapun kelemahan dari metode ini yakni diperlukannya penentuan nilai untuk parameter k (banyaknya tetangga terdekat).

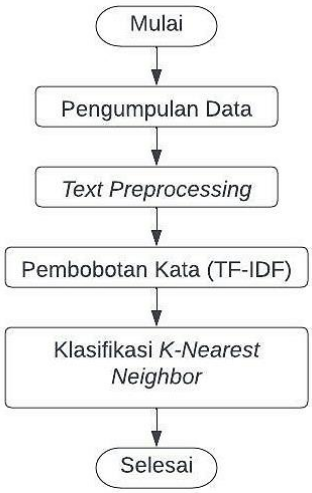
Penelitian mengenai topik klasifikasi berita atau teks menggunakan metode *k-nearet neighbor* terus berkembang. Hal ini dibuktikan dengan banyaknya penelitian mengenai topik tersebut. Seperti penelitian yang dilakukan oleh Fauziah berjudul "Klasifikasi Berita Politik Menggunakan Algoritma K-Nearest Neighbor", diperoleh hasi yang baik ketika nilai K=9 nilai recall , f-measure dan precision sebesar 100%

dengan data latih 210 dan pengujian ke tiga 270 berita [5]. Penelitian berikutnya ialah yang dilakukan oleh Briliansyah dengan judul "Sistem Klasifikasi Kategori Berita Menggunakan Metode K-Nearest Neighbor", berdasarkan uji coba sebanyak 20 data berita diperoleh persentase nilai akurasi sebesar 74%, *f-measure* 35%, *specifity 84%, UAC* 60%. *precision* 35%, *recall* 35% [6]. Selanjutnya ialah penelitian yang dilakukan oleh Sagita, Enri, dan Primajaya berjudul "Klasifikasi Berita Clickbait Menggunakan K-Nearest Neighbor (KNN)", mendapati proses klasifikasi terbaik adalah model dengan nilai k=11, yakni mendapat nilai akurasi sebesar 71% [7].

Penelitian lainnya ialah yang dilaksanakan oleh Palma, Murdiansyah, dan Astuti berjudul "Klasifikasi Teks Artikel Berita Hoaks Covid-19 dengan Menggunakan Algoritma K-Nearest Neighbor", penggunaan k=5 pada penelitian ini lebih akurat dibandi k=3, k=7, dan k=9. Dengan menggunakan perbandingan data latih:data uji sebesar 80%:20% menghasilkan akurasi sebesar 48%. Selain itu ketika menggunakan validasi k-fold cross validation memakai nilai k=5 menghasilkan akurasi sebesar 42% [8]. Selain itu penelitian yang serupa juga dilakukan oleh Indriyanti, Sugianti, dan Karomi berjudul "Peningkatan Akurasi Algoritma KNN dengan Seleksi Fitur Gain Ratio untuk Klasifikasi Penyakit Diabetes Mellitus", peningkatan akurasi diperoleh pada nilai threshold sebesar 0,152 dengan mempertahankan 4 dari 8 atribut [9]. Penelitian lainnya yakni yang dilakukan oleh Setyaningrum dengan judul "Classification of Twitter Contents using Chi-Square and K-Nearest Neighbor Algorithm", dapat ditarik kesimpulan bahwa algoritma KNN dengan fitur Chi-Square mendapatkan hasil akurasi dilihat dari nilai tetangga terdekat k=3 merupakan hasil terbesar dengan nilai akurasi 65% [10]. Penelitian terakhir yang menjadi rujukan dalam penelitian penulis yakni yang dilakukan oleh Ekienabor berjudul "Text Classification Using KNN with Different Feature Selection Methods", dari penelitian dapat disimpulkan bahwa KNN bekerja lebih baik untuk jumlah fitur yang sedikit. Saat menaikkan jumlah nomor fitur makan kinerja KNN akan menurun. Hal ini disebabkan oleh KNN menggunakan Euclidean yang tidak berarti ketika dimensi data meningkat secara signifikan [11].

3. Metode Penelitian

Penelitian yang akan dilaksanakan ini menggunakan salah satu metode penelitian yakni metode kuantitatif. Pada penelitian kuantitatif menekankan tentang pengujian teori berdasarkan variable yang diukur dan melakukan analisa terhadap data dengan prosedur statistik [12]. Dalam penelitian kuantitatif terdapat batasan dalam lingkup penelitian yang membatasi variabel dan populasi yang dipergunakan dalam penelitian. Penelitian kuantitatif dilakukan dengan rancangan tahapan yang terstruktur dan sesuai dengan sistematika penelitian ilmiah. Berikut ialah tahapan yang akan dilaksanakan dalam penelitian ini.



Gambar 2. Alur Metode Penelitian

Berikut penjelasan tahapan penelitian sesuai gambar diatas:

a. Pengumpulan Data

Proses pengumpulan data yang diperlukan pada penelitian ini dilakukan secara manual dari tanggal 30 Januari 2022 sampai dengan 10 Maret 2022 yang dikumpulkan dari beberapa portal berita khusus daerah Riau. Data yang dikumpulkan berjumlah 510 data berita, setelah itu 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*. Di bawah ini ialah

contoh dari *dataset* berita yang telah dikumpulkan dan diberikan label:

Dataset Berita	Kelas
Intensitas curah hujan yang cukup tinggi beberapa bulan terakhir membuat beberapa titik ruas jalan provinsi yang ada di Kabupaten Siak mengalami kerusakan.	Demokrasi
Pemerintah Kota (Pemko) Pekanbaru sedang mengumpulkan data masyarakat miskin positif Covid-19	Kemiskinan
Gubernur Riau Syamsuar mengungkapkan perkebunan kelapa sawit merupakan sektor paling dominan dalam penyerapan tenaga kerja di Provinsi Riau.	Ketenagakerjaan

Tabel 1. Dataset Berita

Perbandingan data latih dan uji pada *dataset* yang digunakan yakni 70% : 30%, 80% : 20% , serta 90% : 10 %. Berikut rincian perbandingan *dataset* yang digunakan.

Data Berita	Kelas	Pembagian data latih dan data uji					
		Latih	Uji	Latih	Uji	Latih	Uji
		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

b. *Text Preprocessing*

Setelah data siap dikumpulkan, selanjutnya data melalui tahap pra proses teks yang sangat penting dalam melakukan proses klasifikasi data. Tujuan dilakukan tahapan pra proses teks ialah untuk membersihkan data dari komponen yang tidak diperlukan, menyeragamkan bentuk kata dan mengurangi volume kata agar mempermudah proses klasifikasi. Berikut tahapan yang akan dilakukan dalam pra proses teks:

1. *Cleaning*
   Tahap *cleaning* dilakukan pembersihan kata yaitu menghilangkan karakter, simbol, atau identitas pengguna yang tidak digunakan seperti ([ ]@#$%^&*), *URL*, angka, dan *emoticon* yang terdapat dalam *dataset.*
2. *Case Folding*
   Tahapan *case folding* digunakan untuk mengganti semua huruf menjadi huruf kecil (*lowercase*).
3. *Tokenizing*
   Tahap tokenisasi (*Tokenizing*) digunakan untuk memecah kalimat menjadi potongan kata yang disebut juga sebagai token (potongan kata tunggal).
4. *Normalisasi*
   Tahapan *normalisasi* merupakan proses pengubahan kata tidak baku atau kata yang ejaannya salah atau tidak tepat menjadi kata baku dengan memakai kamus normalisasi yang dibuat manual berdasarkan pengecekkan data secara manual.
5. *Removal Stopword*
   Tahapan *removal stopword* merupakan tahap yang dipergunakan untuk menghilangkan kata-kata yang tidak penting dari hasil normalisai sebelumnya. Kata-kata yang dihilangkan berupa kata hubung dan kata keterangan yang tidak digunakan dalam proses klasifikasi. 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* adalah proses menganalisa setiap kata dari hasil *removal stopword* sebelumnya untuk diubah menjadi kata dasar. Proses *stemming* akan menghilangkan imbuhan awalan, akhiran, sisipan, dan kombinasi awalan dan akhiran. Pada proses *stemming* ini menggunakan kelas *StemmerFactory* dari *library* Sastrawi.

c. Pembobotan Kata (TF-IDF)

Tahapan ini akan memberi bobot pada setiap kata (*term*) dari seluruh dokumen akan di berikan bobot. Nilai dari munculnya kata pada sebuah dokumen disebut sebagai *term frequency* (TF). Persamaan yang digunakan untuk menghitung nilai *term frequency* dapat dilihat pada rumus 3.1 sebagai berikut.

$$TF(d,t) = F(d,t) \quad (3.1)$$

Keterangan:
$F(d,t) =$ Kemunculan kata t dalam didokumen d

Kemudian dilakukan perhitungan IDF (*inverse document frequency*). Persamaan idf dapat dilihat pada persamaan 3.2 berikut ini..

$$IDF(t) = log(\frac{N}{df(t)}) \quad (3.2)$$

Keterangan:
$df(t) =$ Document frequency dari kata t
N = Total dokumen latih

Term Frequency Inverse Document Frequency (TF-IDF) menjadi tolak ukur perhitungan statistik dalam proses analisa seberapa penting keberadaan sebuah kata pada sebuah dokumen yang dapat diasumsikan sebagai berikut:

$$w_{ij} = tf_{ij} \; x \; idf_{ij} \quad (3.3)$$

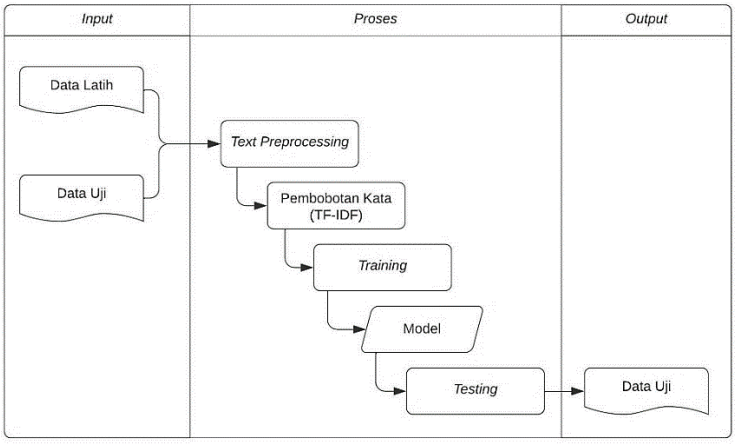
Keterangan
w_{ij} : Bobot *term* terhadap dokumen
tf_{ij} : Jumlah kemunculan kata (*term*) dalam setiap dokumen
idf_{ij} : Jumlah kemunculan kata (*term*) dalam sebuah dokumen

d. Klasifikasi *K-Nearest Neighbor*

Hasil dari pembobotan kata dan perhitungan kemiripan dokumen digunakan pada tahapan analisa k-nearest neighbor. Berikut ialah tahapan umum yang dilakukan pada klasifikasi memakai metode k-nearest neighbor adalah sebagai berikut:

1. Tentukan banyaknya nilai Kyang akan digunakan serta nilai K sebagai pertimbangan dalam penentuan kelas
2. Kemudian hitunglah jarak antara data baru kepada masing-masing data point di dataset.
3. Ambil beberapa K data dengan jarak terdekat, lalu tentukan kelas dari data baru tersebut.

Berikut tahapan proses klasifikasi menggunakan metode *k-nearest neighbor* pada Gambar 3 berikut ini.



Gambar 3. Alur Tahapan Klasifikasi K-Nearest Neighbor

Berikut penjelasan dari Gambar 3 diatas:

1. *Input*

Pada tahapan *input* data yang dipergunakan ialah data berita yang telah dikumpulkan dan diberi label secara manual oleh tim ahli dari bidang sosial Badan Pusat Statistik Provinsi Riau sebanyak 510 berita. 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 *K-Nearest Neighbor*

Pada proses klasifikasi, data akan diproses melalui tahapan *text preprocessing*, pembobotan kata (tf-idf), dan pemodelan *K-Nearest Neighbor*.

3. *Output*

Tahap terakhir yaitu hasil data yang telah diklasifikasikan sesuai kelas demokrasi, kemiskinan, dan ketenagakerjaan.

4. Hasil dan Pembahasan

Berikut merupakan hasil dari setiap proses pada penelitian ini, yaitu klasifikasi berita menggunakan metode *k-nearest neighbor*:

a. *Text Preprocessing*

1. *Cleaning*

Berikut hasil tahapan *cleaning* pada Tabel 3 berikut ini.

Berita sebelum tahapan *cleaning*	**Berita setelah tahapan *cleaning***
Intensitas curah hujan yang cukup tinggi beberapa bulan terakhir membuat beberapa titik ruas jalan provinsi yang ada di Kabupaten Siak mengalami kerusakan.	Intensitas curah hujan yang cukup tinggi beberapa bulan terakhir membuat beberapa titik ruas jalan provinsi yang ada di Kabupaten Siak mengalami kerusakan
Pemerintah Kota (Pemko) Pekanbaru sedang mengumpulkan data masyarakat miskin positif Covid-19	Pemerintah Kota Pemko Pekanbaru sedang mengumpulkan data masyarakat miskin positif Covid 19
Gubernur Riau Syamsuar mengungkapkan perkebunan kelapa sawit merupakan sektor paling dominan dalam penyerapan tenaga kerja di Provinsi Riau.	Gubernur Riau Syamsuar mengungkapkan perkebunan kelapa sawit merupakan sektor paling dominan dalam penyerapan tenaga kerja di Provinsi Riau

Tabel 3. Hasil Cleaning Data

2. *Case Folding*

Berikut hasil tahapan *case folding* pada Tabel 4 berikut ini.

Berita sebelum tahapan *case folding*	**Berita setelah tahapan *case folding***
Intensitas curah hujan yang cukup tinggi beberapa bulan terakhir membuat beberapa titik ruas jalan provinsi yang ada di Kabupaten Siak mengalami kerusakan	intensitas curah hujan yang cukup tinggi beberapa bulan terakhir membuat beberapa titik ruas jalan provinsi yang ada di kabupaten siak mengalami kerusakan
Pemerintah Kota Pemko Pekanbaru sedang mengumpulkan data masyarakat miskin positif Covid 19	pemerintah kota pemko pekanbaru sedang mengumpulkan data masyarakat miskin positif covid 19
Gubernur Riau Syamsuar mengungkapkan perkebunan kelapa sawit merupakan sektor paling dominan dalam penyerapan tenaga kerja di Provinsi Riau	gubernur riau syamsuar mengungkapkan perkebunan kelapa sawit merupakan sektor paling dominan dalam penyerapan tenaga kerja di provinsi riau

Tabel 4. Hasil Case Folding

3. *Tokenizing*

Berikut hasil *tokenizing* pada Tabel 5 berikut ini.

D1	**D2**	**D3**
intensitas	pemerintah	gubernur
curah	kota	riau
hujan	pemko	syamsuar
yang	pekanbaru	mengungkapkan
cukup	sedang	perkebunan
tinggi	mengumpulkan	kelapa
beberapa	data	sawit
bulan	masyarakat	merupakan
terakhir	miskin	sektor
membuat	positif	paling
titik	covid	dominan

ruas jalan provinsi yang ada di kabupaten siak mengalami kerusakan		dalam penyerapan tenaga kerja di provinsi riau

Tabel 5. Hasil Tokenisasi

4. *Normalisasi*
   Berikut hasil *normalisasi* pada Tabel 6 berikut ini

D1	D2	D
intensitas curah hujan yang cukup tinggi beberapa bulan terakhir membuat titik ruas jalan provinsi yang ada di kabupaten siak mengalami kerusakan	pemerintah kota pemko pekanbaru sedang mengumpulkan data masyarakat miskin positif covid	gubernur riau syamsuar mengungkapkan perkebunan kelapa sawit merupakan sektor paling dominan dalam penyerapan tenaga kerja di provinsi riau

Tabel 6. Hasil Normalisasi

5. *Removal Stopword*
   Berikut hasil *removal stopword* pada Tabel 7 berikut ini.

D1	D2	D3
intensitas curah hujan titik ruas jalan provinsi kabupaten siak mengalami kerusakan	pemerintah kota pemko pekanbaru mengumpulkan data masyarakat miskin positif covid	gubernur riau syamsuar perkebunan kelapa sawit sektor dominan penyerapan tenaga kerja provinsi riau

Tabel 7. Hasil Removal Stopword

6. *Stemming*
   Berikut hasil *stemming* pada Tabel 8 berikut ini.

D1	D2	D3
intensitas curah	perintah kota	gubernur riau

hujan	pemko	syamsuar
titik	pekanbaru	kebun
ruas	kumpul	kelapa
jalan	data	sawit
provinsi	masyarakat	sektor
kabupaten	miskin	dominan
siak	positif	serap
alami	covid	tenaga
rusak	isolasi	kerja
	fasilitas	provinsi
	perintah	riau
	opname	

Tabel 8. Hasil Stemming

b. Pembobotan Kata (TF-IDF)
Berikut hasil perhitungan TF-IDF pada Tabel 9 berikut ini.

Term	TF			DF	IDF	TF-IDF		
	D1	D2	D3			D1	D2	D3
alami	1	0	0	1	ln(4/2)+1 = 2,098	2098	0	0
covid	0	1	0	1	ln(4/2)+1 = 2,098	0	2,098	0
curah	1	0	0	1	ln(4/2)+1 = 2,098	2,098	0	0
data	0	1	0	1	ln(4/2)+1 = 2,098	0	2,098	0
dominan	0	0	1	1	ln(4/2)+1 = 2,098	0	0	2,098
gubernur	0	0	1	1	ln(4/2)+1 = 2,098	2,098	0	0
intensitas	1	0	0	1	ln(4/2)+1 = 2,098	2,098	0	0
jalan	1	0	0	1	ln(4/2)+1 = 2,098	2,098	0	0
kabupaten	1	0	0	1	ln(4/2)+1 = 2,098	2,098	0	0
kebun	0	0	1	1	ln(4/2)+1 = 2,098	0	0	2,098
kelapa	0	0	1	1	ln(4/2)+1 = 2,098	0	0	2,098
kerja	0	0	1	1	ln(4/2)+1 = 2,098	0	0	2,098
kota	0	1	0	1	ln(4/2)+1 = 2,098	0	2,098	0
kumpul	0	1	0	1	ln(4/2)+1 = 2,098	0	2,098	0
masyarakat	0	1	0	1	ln(4/2)+1 = 2,098	0	2,098	0
miskin	0	1	0	1	ln(4/2)+1 = 2,098	0	2,098	0
pekanbaru	0	1	0	1	ln(4/2)+1 = 2,098	0	2,098	0
pemko	0	1	0	1	ln(4/2)+1 = 2,098	0	2,098	0
positif	0	1	0	1	ln(4/2)+1 = 2,098	0	2,098	0
Provinsi	1	0	1	2	ln(4/2)+1 = 1,405	1,405	0	1,405
perintah	0	1	0	1	ln(4/2)+1 = 2,098	0	2,098	0
Riau	0	0	2	2	ln(4/2)+1 = 4,197	0	0	4,197
ruas	1	0	0	1	ln(4/2)+1 = 2,098	2,098	0	0
rusak	1	0	0	1	ln(4/2)+1 = 2,098	2,098	0	0
sawit	0	0	1	1	ln(4/2)+1 = 2,098	0	0	2,098
sektor	0	0	1	1	ln(4/2)+1 = 2,098	0	0	2,098
serap	0	0	1	1	ln(4/2)+1 = 2,098	0	0	2,098
siak	1	0	0	1	ln(4/2)+1 = 2,098	2,098	0	0
syamsuar	0	0	1	1	ln(4/2)+1 = 2,098	0	0	2,098
tenaga	0	0	1	1	ln(4/2)+1 = 2,098	0	0	2,098

Term	TF			DF	IDF	TF-IDF		
	D1	D2	D3			D1	D2	D3
titik	1	0	0	1	ln(4/2)+1 = 2,098	2,098	0	0

Tabel 9. Hasil TF-IDF

c. Klasifikasi *K-Nearest Neighbor*

Data berita yang sudah di praproses dan diberikan bobot pada kata (tf-idf) selanjutnya akan melalui proses klasifikasi menggunakan metode *k-nearest neighbor* agar menghasilkan model yang dapat digunakan untuk proses klasifikasi data berita baru tanpa pemberian label manual. Pada tahapan ini proses pembelajaran dan pelatihan data dilakukan. Proses pembelajaran ini dilakukan dengan melatih model menggunakan dataset yang sudah diberi label. Pada proses pelatihan data dilakukan uji coba menggunakan data yang belum diberikan label. Hasil dari perhitungan model KNN pada penelitian ini perlu ditingkatkan. Salah satu cara peningkatan akurasi dalam model KNN ini ialah memakai perhitungan *cosine similarity.*

d. Pengujian *Confusion Matrix*

Pengujian *confusion matrix* dihitung berdasarkan pembagian data latih dan data uji dari data berita yang berjumlah 510 berita. Pembagian data dalam penelitian ini menggunakan perbandingan 3 skenario yakni 70% data latih : 30% data uji, 80% data latih : 20% data uji, dan 90% data latih: 10 % data uji dengan nilai k = 3.

1. Pengujian terhadap 70% data latih dan 30% data uji

Berikut hasil pengujian *confusion matrix* untuk 70% data latih dan 30% data uji pada Tabel 10 berikut ini.

Predicted Classification	*Actual Classification*		
	Demokrasi Positive	Kemiskinan Positive	Ketenagakerjaan Positive
Demokrasi Negative	44	5	9
Kemiskinan Negative	1	44	2
Ketenagakerjaan Negative	5	4	39

Tabel 10. Hasil Confusion Matrix Pembagian Data 70%:30%

$$\text{Perhitungan akurasi} = \frac{44+44+39}{44+5+9+1+44+2+5+4+39} \; x\; 100\%$$
$$= \frac{127}{153} \; x\; 100\%$$
$$= 83\%$$

2. Pengujian terhadap 80% data latih dan 20% data uji

Berikut hasil pengujian *confusion matrix* untuk 80% data latih dan 20% data uji pada Tabel 11 berikut ini.

Predicted Classification	*Actual Classification*		
	Demokrasi Positive	Kemiskinan Positive	Ketenagakerjaan Positive
Demokrasi Negative	30	2	5
Kemiskinan Negative	2	33	1
Ketenagakerjaan Negative	1	3	26

Tabel 11. Hasil Confusion Matrix Pembagian Data 80%:20%

$$\text{Perhitungan akurasi} = \frac{30+33+26}{30+2+5+2+33+1+1+3+26} \; x\; 100\%$$
$$= \frac{89}{102} \; x\; 100\%$$
$$= 87\%$$

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

Berikut hasil pengujian *confusion matrix* untuk 90% data latih dan 10% data uji pada Tabel 12 berikut ini.

Predicted Classification	*Actual Classification*		
	Demokrasi Positive	Kemiskinan Positive	Ketenagakerjaan Positive
Demokrasi Negative	13	1	5
Kemiskinan Negative	2	15	2
Ketenagakerjaan	0	2	11

Negative			

Tabel 12. Hasil Confusion Matrix Pembagian Data 90%:10%

Perhitungan akurasi $= \frac{13+15+11}{13+1+5+2+15+2+0+2+11} \; x \; 100\%$

$= \frac{39}{51} \; x \; 100\%$

$= 76\%$

Untuk hasil akurasi pengujian lengkap pada tiap pembagian data menggunakan 7 nilai K dapat dilihat melalui tabel 13 berikut ini.

Nilai K	Akurasi (%)		
	Split Data 70%:30%	*Split Data* 80%:20%	*Split Data* 90%:10%
3	83%	87%	76%
5	83%	85%	74%
7	78%	80%	74%
9	79%	84%	74%
11	79%	85%	78%
13	81%	83%	76%
15	79%	82%	78%

Tabel 13. Hasil Uji Akurasi

5. Kesimpulan

Berdasarkan hasil dari penelitian yang telah dilaksanakan, dapat diambil beberapa kesimpulan sebagai berikut:

1. Metode klasifikasi *k nearest neighbor* terbukti dapat digunakan dalam proses klasifikasi berita.
2. Nilai akurasi paling tinggi yang diperoleh dari proses klasifikasi yaitu 87% dengan nilai k=3 dan pembagian data latih 80% dan data uji 20% dari *dataset* yang digunakan.

Daftar Pustaka

[1] Y. Pusparisa, "Masyarakat Indonesia Paling Banyak Akses Berita dari Media Daring," *databoks*, 2021. .
[2] S. N. Asiyah and K. Fithriasari, "132747-ID-klasifikasi-berita-online-menggunakan-me," *J. Sains dan Seni ITS*, vol. 5, no. 2, pp. 317–322, 2016.
[3] A. Y. Muniar, P. Pasnur, and K. R. Lestari, "Penerapan Algoritma K-Nearest Neighbor pada Pengklasifikasian Dokumen Berita Online," *Inspir. J. Teknol. Inf. dan Komun.*, vol. 10, no. 2, p. 137, 2020.
[4] R. N. Devita *et al.*, "PERBANDINGAN KINERJA METODE NAIVE BAYES DAN K-NEAREST NEIGHBOR UNTUK KLASIFIKASI ARTIKEL BERBAHASA INDONESIA PERFORMANCE COMPARISON OF NAIVE BAYES AND K-NEAREST NEIGHBOR," vol. 5, no. 4, pp. 427–434, 2018.
[5] D. A. Fauziah, A. Maududie, and I. Nuritha, "Klasifikasi Berita Politik Menggunakan Algoritma K-nearst Neighbor," *Berk. Sainstek*, vol. 6, no. 2, p. 106, 2018.
[6] F. Briliansyah, "Sistem klasifikasi kategori berita menggunakan metode k-nearest neighbor," 2020.
[7] R. Sagita, U. Enri, and A. Primajaya, "Klasifikasi Berita Clickbait Menggunakan K-Nearest Neighbor (KNN)," vol. 5, no. 2, pp. 230–238, 2020.
[8] B. K. Palma, D. T. Murdiansyah, and W. Astuti, "Klasifikasi Teks Artikel Berita Hoaks Covid-19 dengan Menggunakan Algotrima K- Nearest Neighbor," vol. 8, no. 5, pp. 10637–10649, 2021.
[9] Indriyanti, D. Sugianti, and M. A. Al Karomi, "Peningkatan Akurasi Algoritma KNN dengan Seleksi Fitur G ain Ratio untuk Klasifikasi Penyakit Diabetes Mellitus," *IC-Tech*, vol. 7, no. 2, pp. 1–6, 2017.
[10] Y. D. Setiyaningrum, A. F. Herdajanti, C. Supriyanto, and Muljono, "Classification of twitter contents using chi-square and K-nearest neighbour algorithm," *Proc. - 2019 Int. Semin. Appl. Technol. Inf. Commun. Ind. 4.0 Retrosp. Prospect. Challenges, iSemantic 2019*, pp. 78–81, 2019.
[11] E. Ekienabor, "International Journal of Research Publications Volume-9 , Issue-1 , July 2018," *Int. J. Res. Publ.*, vol. Volume 8, no. July, 2018.
[12] C. R. Semiawan, *Metode Penelitian Kuantitatif*. 2017.