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

Klasifikasi Berita Menggunakan Metode *Support Vector Machine*

Robbi Nanda[1], Elin Haerani[2], Siska Kurnia Gusti[3], Siti Ramadhani[4]
[1,2,3,4]Teknik Informatika Universitas Islam Negeri Sultan Syarif Kasim Riau
Jl. H.R Soebrantas No. 155 KM. 18 Simpang Baru, Pekanbaru 28293
Corresponding author's email : *11850114572@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 **-** Berita adalah sebuah informasi mengenai peristiwa yang terjadi di suatu lokasi yang bisa disajikan dalam bentuk teks maupun visual. Berita bisa ditemukan di berbagai portal berita dan media cetak. Umumnya setiap berita dikelompokan berdasarkan kategori umum seperti ekonomi, politik, olahraga, dll. Permasalahan yang muncul adalah bagaimana cara untuk melakukan pengelompokan pada data berita yang biasanya berjumlah hingga ribuan karakter kedalam kategori yang lebih spesifik. Permasalah ini dapat diselesaikan dengan cara menerapkan text mining dengan memanfatakan algoritma klasifikasi untuk mendapatkan sebuah model fungsi yang merepresentasikan tiap kategori berita. Salah satu algoritma klasifikasi yang cukup tangguh untuk melakukan proses klasifikasi teks adalah *Support Vector Machine.* Penelitian ini menggunakan 510 data berita dengan batasan klasifikasi 3 kategori berita. Algoritma SVM mendapatkan hasil akurasi tertinggi di 88% untuk nilai parameter C =1, kernel Linear dengan pembagian data uji dan data latih sebesar 90% dan 10 %.
Kata kunci **:** *Berita, Klasifikasi, Support Vector Machine, Text Mining*

Abstract **-** News is information about events that occur in a location that can be presented in text or visual form. News can be found on various news portals and print media.Generally each news is grouped by general categories such as economics, politics, sports, etc. The problem is how to group news data into more specific categories.This problem can be solved by applying text mining using the classification algorithm to obtain a function model that represents each news category. One of the classification algorithms that is strong enough to do the text classification process is the Support Vector Machine. This study uses 510 news sample with a classification limit of 3 news categories. The SVM algorithm gets the highest accuracy at 88% for the parameter value C = 1, and Linear kernel with the distribution of test data and training data is 90% and 10%.
Keywords ***:*** *Classification, News, Support Vector Machine, Text Mining*

1. Pendahuluan

Indonesia saat ini telah berada dalam masa perkembangan teknologi informasi. Akses internet yang semakin mudah, juga mendorong kemajuan teknologi di berbagai bidang. Kemajuan teknologi telah berdampak pada akses informasi lebih cepat dan mudah. Dari sumber yang di peroleh Kementerian Komunikasi dan Informatika diketahui bahwa saat ini pengguna internet di Indonesia telah melebihi delapan puluh juta orang. Angka ini menempatkan indonesia di peringkat ke delapan untuk pengguna internet terbesar di Dunia[1]. Dengan adanya ledakan penggunaan internet tersebut, tentu berdampak pada peningkatan penggunaan media internet sebagai sarana informasi khususnya berita online.

Berita pada umumnya akan disampaikan dengan beberapa kategori seperti ekonomi, teknologi, olahraga, kesehatan, dan lain-lainyna. Dalam penelitian ini, berita online digunakan sebagai bahan penelitian yang dilakukan di BPS Prov Riau. BPS atau Badan Pusat Statistik menggunakan berita berita yang terjadi setiap harinya untuk di olah menjadi bahan penunjang hasil statistik peristiwa di provinsi riau. Berita ini nantinya akan diambil kemudian di klasifikasikan kedalam kategori yang lebih spesifik dan kemudian dilakukan scoring untuk tiap berita tersebut.

Sejauh ini pengelompokan berita kedalam berbagai kategori tersebut dilakukan secara manual oleh staff di BPS Prov. Riau. Dalam prosesnya, berita akan dibaca kemudian ditafsirkan untuk mengklasifikasikan berita kedalam kategori yang tepat. Permasalahan lain yang timbul dalam pengelompokan teks secara manual adalah kemiripian isi yang membutuhkan ketelitian lebih untuk mengklasifikasikan teks berita sesuai dengan isinya. Dari permasalahan tersebut, diperlukan sebuah alat atau *tools* yang dapat melakukan klasifikasi atau pengelompokan secara otomatis. Metode yang tepat untuk mengatasi permasalahan ini adalah menggunakan *text mining.*

Text Mining adalah sebuah proses terstruktur untuk mencari atau melakukan pembentukan token *text* terstruktur rapi dan penambangan informasi penting dari data teks [2]. Text mining adalah langkah dari pemahaman teks yang dilakukan oleh komputer secara otomatis untuk mencari sebuah informasi atau rangkaian pola penting dari sebuah teks [3].

Dalam penelitian ini metode klasifikasi yang di implementasikan adalah sebuah algoritma pembelajaran yang cukup populer yaitu *algoritma Support Vector Machine (SVM).* Dalam beberapa studi literatur SVM memiliki tingkat akurasi paling tinggi yaitu sebesar 88% dalam mempelajari klasifikasi teks [4]. Apabila

membandingkan dengan algoritme *Naive Bayes Classifier* yang mana ini juga merupakan algoritma yang cukup populer dalam pengklasifikasian teks, algoritma SVM cukup tangguh untuk memproses data teks. Menurut Christianini dan J Shawe-Taylor dikutip dari [4] *Support Vector Machine (SVM)* adlaah sebuah sistem pembelajaran yang mempelajari pola dengan memahami ruang hipotesis yang berupaya untuk membentuk fungsi - fungsi linear dalam sebuah ruang fitur *(feature space)* berdimensi tinggi. Algortima ini memanfatkan learning bias dan optimasi untuk mendapatkan hasil yang baik. Tujuan dari penelitian ini yaitu untuk melakukan klasifikasi berita menggunakan metode algritma klasifikasi *Support Vector Machine* yang nantinya hasil model ini akan digunakan sebagai pendukung kegiatan survei bidang sosial di BPS Prov. Riau.

2. Tinjauan Pustaka

2.1. *Text Mining*

Text Mining merupakan bagian dari data mining yang sama sama memiliki tujuan untuk menambang data untuk mencari sebuah pola atau keterkaitan unik yang merupakan perwakilan isi atau ciri khas sebuah dokumen teks. taAda banyak tahapan dalam *text mining*[5]. Berikut adalah beberapa tahapan umum pada proses Text Mining menurut Nugroho 2011 dikutip dari [6], yaitu :

1. Text Preprocessing, yaitu pengolahan text agar siap untuk dijadikan data dan tidak mengalami masalah pada proses setelahnya.
2. Case Folding, adalah mengubah karakter *uppercase* menjadi *lowercase*. Hal ini dilakukan karena banyak bahasa pemograman yang sensitif terhadap huruf besar dan huruf kecil.
3. Filtering, adalah proses penyaringan untk mengambil kata kata yang dibutuhkan.
4. Tokenizing, yaitu tahapan untuk memecah sebuah kalimat menjadi kata per kata. Hasil tokenizing umumnya dipisahkan dengan sebuah karakter atau whitespace.
5. Text Transformation, pengubahan teks untuk masuk pada representasi dokumen yang diinginkan.

2.2. Klasifikasi

Proses klasifikasi adalah proses mendapatkan fungsi atau model yang menjelaskan atau membedakan konsep/kelas data, ketika model telah ditemukan makan model ini dapat digukanakan untuk melakukan prediksi kelas data baru. Tujuan dari klasifikasi adalah untuk memberikan kelas kepada beberapa data sesuai dengan model yang telah dipelajari. Pengkelasan ini dapat masuk ke berbagai kelas tidak hanya terbatas pada satu kelas saja. Adapun beberapa tahap yang umum dilakukan adalah sebagai berikut :

1. Tahap Pra-proses (preprocessing), yaitu tahap awal dari penelitian ini dan merupakan tahap yang sangat penting agar pemrosesan data dapat dilanjutkan.
2. Pembobotan. Yaitu proses untuk memberikan bobot pada data, hal ini umum dilakukan terhadap data berbentuk teks.
3. Proses Klasifikasi. Yaitu proeses yang dilakukan setelah semua data bersih dan memiliki bobot yang dapat dihitung menggunakan bebrapa metode.

Data berita berupa teks yang sangat panjang, untuk itu dalam penelitian ini sangat penting utnuk dilakukan tahapan data *preprocessing*. Hal ini bertujuan untuk mengurangi kata kata yang tidak diperlukan dalam proses klasifikasi dan meningkatkan akurasi dari proses pengklasifikasian

2.3. Data *Preprocessing*

Tujuan dari tahapan pre processing ini yaitu membersihkan data dari faktor yang tidak diperlukan sehingga proses penambangan data menjadi lebih cepat dan akurat. Adapun tahapan *preprocessing* text dikutip dari [7] adalah sebagai berikut:

1. Cleaning, adalah sebuah tahapan untuk memberihkan data dari faktor faktor tidak penting seperti kesalahan bahasa, kata hubung, tanda baca, dan beberapa teks yang tidak dibutuhkan.
2. Case Folding, adalah sebuah kegiatan mentransformasi kata yang memiliki huruf besar menjadi huruf kecil.
3. Tokenizing, adalah tahapan untuk membagi atau memecah kata yang awalnya dalam bentuk kalimat utuh menjadi penggalan kata perkata. Ini dilakukan agar proses pengolahan teks menjadi lebih mudah dan dapat meningkatkan akurasi hasil prediksi.
4. Stopwords, merupakan kata yang tidak bersifat unik dan menggambarkan karakteristik sebuah data yang diolah. Dalam proses ini data akan dihapus dari kata kata yang tidak dibutuhkan tersebut. Selain dapat digunakan kamus stopword, dapat juga dengan menambahkan kamus kata yang dibuat sendiri.
5. Stemming, dalam tahapan ini setiap kata akan dihapus dari imbuhan yang di milikinya. Imbuhan ini bisa saja berada di awal kata atau berada di akhir kata sehingga hasilnya akan menjadi kata dasar dari kata tersebut.

2.4.Pembobotan Kata (TF-IDF)

Term Frequency atau (TF) , yaitu bobot term pada suatu data dalam hal ini data berita, yang diperhitungkan berdasarkan *n* atau tingkat kemunculannya dalam tumpukan data. TF merupakan jumlah dari banyaknya kata dalam satu dokumen, semakin besar nilai TF dari seuatu dokumen maka akan semakin tinggi bobot dokumen nya. *Inverse Document Frequency* (IDF) *factor*, yaitu factor untuk mengurangi nilai dari TF sebelumnya. Jika

sebuah kata sering muncul dalam dokumen tersebut makan nilai TF atau bobotnya akan semakin kecil. Semakin jarang suatu kata muncul atau faktor *(term scarcity)* makan bobot dari TF-IDF akan semakin besar. Hal ini karena semakin jarang sebuah kata dalam dokumen lain, akan menunkukan bahwa kata tersebut mempunyai pengaruh yang besar. Metode TF-IDF adalah metode yang paling sering dipakai untuk mencari bobot dokumen.

Pembobotan kata atau yang disebut *(term weighting)* menghitung dimensi dari *feature vector* berdasarkan nilai (IDF) *Inverse Document Frequency* nya. *Term weighting* membentuk formula sebagai berikut:

$$tf * idf \tag{1}$$

- tf = tingkat kemunculan kata dari suatu data teks
- idf = inverse document frequency

Metode *tf*idf* ini dipilih karena diketahui bahwa dengan mencari dimensi dari *feature vector* dengan (IDF) *Inverse Document Frequency* nya dapat meningkatkan akurasi dan performa pencarian model[8]. Kata-kata yang muncul hanya dalam sedikit dokumen biasanya lebih bernilai daripada yang muncul di banyak dokumen. IDF dari sebuah kata dasar w_i dapat dirumuskan sebagai berikut :

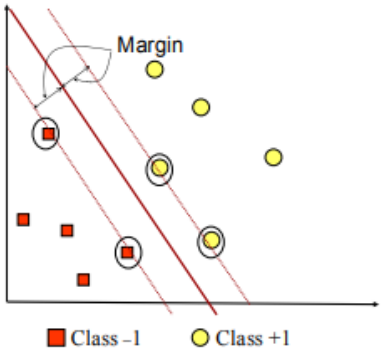
$$IDF(w_i) = log\ (n/DF(w_i)) \tag{2}$$

- n = jumlah total dokumen dalam data
- w_i = kata dasar *(word stem)*
- DF(w_i) = jumlah dokumen dimana kata w_i muncul
-

IDF akan rendah jika sebuah kata muncul di banyak dokumen dan tinggi jika kata hanya muncul dalam satu dokumen. Panjang dari dokumen juga mempengaruhi IDF. Sebuah kata yang muncul di sebuah dokumen pendek lebih berharga daripada yang muncul pada dokumen panjang.

2.5. *Support Vector Machine (SVM)*

Data yang bisa dipisahkan secara fungsi linier disebut dengan linearly separable data. Contohnya $\mathbf{x_i} = \{x_i,\ldots, x_n\}$, $\mathbf{x_i} \in \Re^n$ adalah data set dan $\mathbf{y_i} \in \{+1,-1\}$ adalah label dari dataset x_i . Ada berbagai alternatif bidang pemisah dilihat dari gambar 2.1 yang dapat membagi semua data berdasarkan dengan kelasnya masing masing. Bisa memisahkan data tidak cukup untuk menjadi bidang pemisah terbaik, faktor margin yang paling besar juga merupakan faktor pentin untuk mencari bidang pemisah terbaik. [9].



Gambar 1. Hyperplane pada SVM Linear

Support vector adalah point data yang berada dalam bidang pembatas. Jika diperhatikan pada gambar 1 diatas terlihat bahwa sepasang bidang sejajar memisahkan dua buah kelas yang berbeda. Setiap bidang memisahkan kelas yang berbeda. Dalam perhitungan matematika dapat dibentuk persamaan seperti berikut:

$$\begin{aligned} \boldsymbol{x_i}\mathbf{w} + \mathbf{b} &\geq +1, y_i = +1 \\ \boldsymbol{x_i}\mathbf{w} + \mathbf{b} &\leq -1, y_i = -1 \end{aligned} \tag{3}$$

$i = 1,2, \ldots, p$

Variabel **w** adalah bidang normal sedangkan b adalah posisi bidang cadangan dihitung center koordinat. Perhitungan margin (jarak) antara bidang pembatas (berdasarkan rumus jarak garis ke titik pusat) adalah :

$$\frac{1-b-(-1-b)}{||\boldsymbol{W}||} = \frac{2}{||\boldsymbol{W}||}.$$

Nilai margin ini dapat dimaksimalkan namun harus tetap sesuai dengan rumus (3). Perkalian antara b dan **w** dengan sebuah konstanta, dapat menghasilkan sebuah nilai yang dikalikan dengan konstata yang sama. Oleh karena itu, jika diperhatikan pada persamaan diatasmerupakan batasan skala yang dapat memenuhi dengan melakukan penskalaan ulang pada b dan **w**. Jika memaksimalkan $\frac{1}{||\boldsymbol{W}||}$ sama dengan meminimalkan $||\mathbf{w}||^2$ dan jika masing masing bidang pembatas pada persamaan (3) dapat dibentuk dalam pertidaksamaan (4) berikut:

$$y_i(\boldsymbol{x_i w} + \boldsymbol{b}) - \mathbf{1} \geq \mathbf{0} \tag{4}$$

maka untuk mencari *dividing field* terbaik dengan value margin terbesar dapat di bentuk formula menjadi masalah optimasi konstrain sebagai berikut:

$$\begin{aligned} &\min \frac{1}{2} \left|\left|\boldsymbol{w}\right|\right|^{\mathbf{2}} \\ &dengan\, y_i(\boldsymbol{x}_i w + b) - 1 \geq 0 \end{aligned} \tag{5}$$

Dalam kenyataanya tidak semua permasalahan dapat diselesaikan secara linear. Metode tambahan untuk menyelesaikan masalah tersebut adalah menggunakan metode kernel. Dengan penggonaan meteode kernel ini, sebuah data akan di transformasikan kedalam bentuk dimensi yang lebih tinggi fitur *(feature space)* yang nantinya data bisa dibagi secara linier pada feature space. Fungsi kernel untuk menjadi sebuah fungsi perlu memenuhi teorema Mercer yang menyatakan bahwa hasil darimatriks kernel harus setengah menuju pasti positif. Ada beberapa fungsi kernel yang umum dipakai pada metode SVM dikutip dari[7] yaitu :

Kernel Linier	$K(x_i, x) = x_i^T x$
Kernel Polynomial	$K(x_i, x) = \left(\gamma x_i^T x + r\right)^p, \gamma > 0$
Kernel Radial Basis Function (RBF)	$K(x_i, x) = exp(-\gamma \parallel x - x_i \parallel^2)$
Sigmoid Kernel	$K(x_i, x) = tanh\left(\gamma x_i^T x + r\right)$

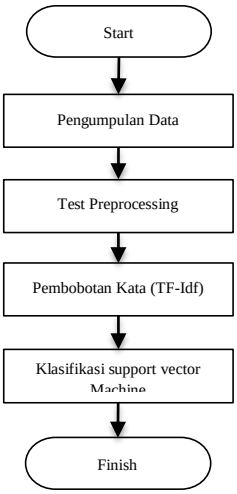
Tabel 1. Fungsi Kernel

2.5. Penelitian Terkait

Saat ini sudah ada banyak penelitian yang menerapkan algoritma *Support Vector Machine* dalam melakukan klasifikasi khususnya klasifikasi Text. Umumnya algoritma ini menghasilkan akurasi yang cukup tinggi dengan rata rata berada diatas 80%. Berikut beberapa penelitian terdahulu untuk klasifikasi text menggunakan *Support Vector Machine*. Penelitian yang dilakukan Oleh Siti Nur Asiyah[7], dengan memanfaatkan algoritma SVM dan K Nearest Neighbor mendapatkan nilai akurasi di 93,2% untuk algortima SVM. Dalam penelitian ini dilakukan klasifikasi berita online kedalam beberapa kategori. Kesimpulan yang didapatkan SVM terbukti memiliki akurasi yang lebih baik jika dibandingkan dengan algoritma K-Nearest Neighbor. Selain itu penelitian yang dilakukan Khrisna Dini Yunita Sari [8] menyimpulkan bahwa SVM bekerja dengan baik untuk mengatasi data dalam jumlah besar dan berdimensi besar. Penelitian oleh Nurfadillah [10] yaitu klasifikasi Topik Tweet mendapatkan hasil akurasi sebesar 96.2% menggunakan algoritma SVM. Selain itu menurut penlitian Fazlur[11] dalam melakukan klasifikasi data tweet menggunakan algoritma SVM kernel RBF didapatkan hasil akurasi sebesar 95%.

3. Metode penelitian

Metode Penelitian yang dilakukan dalam penelitian ini adalah metode kuantitatif. Dalam penelitian kuantitatif terdapatbatasan 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. Berikut tahapan yang akan dilakukan dalam penelitian ini.



Gambar 2. Tahapan Penelitian

3.1. Sumber Data

Dalam proses pengumpulan data, pengambilan data dilakukan dari tanggal 14 Februari sampai dengan 16 Maret 2022 yang diambil dari beberapa portal berita Riau seperti Riaupos, Tribun Pekanbaru, Go Riau. Proses pengambilan berita ini dilakukan dengan cara menyalin secara manual isi konten berita dari tiap portal berita. Data berita yang digunakan adalah isi teks berita secara keseluruhan. Perbandingan data yang akan digunakan yaitu 70% data latih dengan 30% data uji, 80% data latih dengan 20% data uji, dan terakhir 90% data latih dengan 10% data uji. Adapun skenario pembagian data lebih jelas sebagai berikut :

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

3.2. Pelabelan Manual

Tahapan selanjutnya setelah data berhasil dikumpulkan yaitu proses pelabelan manual. Pelabelan ini bertujuan untuk memberikan label dari tiap berita yang sesuai kedalam tiga kelas yang telah di tentukan, yaitu Demokrasi, Ketenagakerjaan, dan Kemiskinan. Proses pelabelan ini akan dilakukan oleh penulis dengan di bimbing oleh staff ahli di BPS Provinsi Riau yang bertanggung jawab dalam bidang pengolahan data indeks di provinsi Riau. Berikut adalah contoh dari data berita yang telah dikumpulkan dan diberi label kelas secara manual.

No.	Berita	Label
1.	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
2.	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
3.	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 3. Pelabelan manual

3.3. *Text Preprocessing*

Setelah proses pelabelan selesai, maka proses selanjutnya yaitu *Preprocessing*. Tahap *Preprocessing* ini adalah tahap yang sangat penting dalam melakukan mining data terutama untuk data berbentuk teks. Dalam data teks ada banyak atribut atau value yang tidak berguna dan akan menghambat proses klasifikasi. Ada beberapa tahapan *preprocessing* yang akan dilakukan dalam penelitian ini, yaitu :

1. *Cleaning*
   Cleaning atau pembersihan data ini dilakukan dengan cara membersihkan data seperti penghilangan karakter atau simbol dan identitas pengguna yang tidak diperlukan seperti URL, emoticon, tanda simbol dan tanda baca
2. *Case folding*
   Pada tahap case folding dilakukan pengubahan semua huruf besar menjadi huruf kecil (lowercase). Pengubahan ini dilakukan untuk memudahkan komputasi karena umumnya bahasa pemograman sangat sensitif terhadap huruf besar dan huruf kecil.
3. *Tokenizing*
   Tokenizing adalah proses untuk membentuk kata perkata dari yang awal nya berbentuk kalimat. Umumnya hasil pemenggalan ini di pisahkan dengan tanda strip atau *whitespace*.
4. *Normalisasi*
   Sebuah data teks bisa saja tidak normal. Tidak normal dalam artian penggunaan basa atau penggunaan kata yang tidak tepat.
5. *Remove Stopword*
   Remove stopword adalah proses untuk menghilangkan kata yang dirasa tidak berguna pada proses klasifikasi. Kata kata yang tidak berguna ini bisa seperti kata kata yang tidak menggambarkan ciri sebuah dokumen ataupun kata kata yang bersifat umum seperti kata hubung. Proses ini memanfaatkan kamus stopword yang bisa dibuat sendiri atau mengunduh kamus stopword yang telah ada di internet.
6. *Stemming*
   Pada tahapan stemming ini data akan diubah kedalam bentuk kata dasarnya, hal ini bisa dilakukan dengan cara menghilangkan awalan, akhiran, sisipan, dan confixes (kombinasi awalan dan akhiran).

3.4. Pembobotan (TF-IDF)

Term frequency – inverse document frequency merupakan teknik untuk mencari hubungan kata perkata pada sebuah dokumen dan hubungannya ke label dengan cara memberikan bobot pada tiap kata *(term)*. *Term* diambil dari hasil pada proses *preprocessing remove stopword*. *Term frequency* berfungsi untuk menghitung jumlah keberadaaan term dalam satu dokumen, sedangkan *document frequency* berfungsi untuk melakukan perhitungan seberapa banyak suatu kata muncuk dalam dokumen lain.

3.5 Klasifikasi *Support Vector Machine*

Pada tahap ini kita melakukan pembelajaran dan uji data. Pengujian pada model dilakukan dengan menggunakan data testing untuk mengetahui nilai akurasi dan klasifikasinya. Hasil akurasi dari metode *support vector machine* tidak bisa langsung digunakan. Karena dalam SVM ada banyak faktor yang bisa meningkatkan hasil akurasi dengan cara melakukan cek performa dari berbagai faktor. Untuk mencari performansi yang baik perlu dilakukan *Hyperparameter tunning* atau pencarian paramete terbaik. Salah satu metode pencarian terbaik yaitu *Grid Search cross validation* yang berfungsi untuk menentukan nilai parameter C dan parameter kernel yang tidak overfit data pelatihan.

4. Hasil dan pembahasan

Berikut merupakan hasil dari setiap proses dalam penelitian ini, yaitu klasifikasi beritamenggunakan metode *naïve bayes classifier*:

4.1. *Text Preprocessing*

1. *Cleaning*

Berikut hasil tahapan *cleaning* pada Tabel 4 berikut.

Berita sebelum tahapan *cleaning*	**Berita setelah tahapan *cleaning***
Rapat Paripurrna dengan agenda pelantikan Pergantian Antar Waktu (PAW) Wakil Ketua DPRD	Rapat Paripurrna dengan agenda pelantikan Pergantian Antar Waktu PAW Wakil Ketua DPRD
Pemerintah Kota (Pemko) Pekanbaru sedang mengumpulkan data masyarakat miskin positif Covid-19	Pemerintah Kota Pemko Pekanbaru sedang mengumpulkan data masyarakat miskin positif Covid
Himbauan kepada Badan Operasi Bersama (BOB) PT Bumi Siak Pusako (BSP)-Pertamina Hulu, agar menunda pengurangan tenaga kerja.	Himbauan kepada Badan Operasi Bersama BOB PT Bumi Siak Pusako BSP Pertamina Hulu agar menunda pengurangan tenaga kerja

Tabel 4. Hasil Cleaning Data

2. *Case Folding*

Berita sebelum tahapan *case folding*	Berita setelah tahapan *case folding*
Rapat Paripurrna dengan agenda pelantikan Pergantian Antar Waktu PAW Wakil Ketua DPRD	rapat paripurrna dengan agenda pelantikan pergantian antar waktu paw wakil ketua dprd
Pemerintah Kota Pemko Pekanbaru sedang mengumpulkan data masyarakat miskin positif Covid	pemerintah kota pemko pekanbaru sedang mengumpulkan data masyarakat miskin positif covid
Himbauan kepada Badan Operasi Bersama BOB PT Bumi Siak Pusako BSP Pertamina Hulu agar menunda pengurangan tenaga kerja	himbauan kepada badan operasi bersama bob pt bumi siak pusako bsp pertamina hulu agar menunda pengurangan tenaga kerja

Tabel 5. Hasil Case Folding

3. *Tokenizing*

Berikut hasil *tokenizing* pada Tabel 5 berikut :

D1	D2	D3
Rapat paripurrna dengan agenda pelantikan pergantian antar waktu paw wakil ketua dprd	pemerintah kota pemko pekanbaru sedang mengumpulkan data masyarakat miskin positif covid	himbauan kepada badan operasi bersama bob pt bumi siak pusako bsp pertamina hulu agar menunda pengurangan tenaga kerja

Tabel 5. Hasil tokenizing

4. *Normalisasi*

Berikut hasil *normalisasi* pada Tabel 6 berikut.

D1	D2	D3
Rapat paripurrna dengan agenda pelantikan pergantian antar waktu paw wakil ketua dprd	pemerintah kota pemko pekanbaru sedang mengumpulkan data masyarakat miskin positif covid	himbauan kepada badan operasi bersama bob pt bumi siak pusako bsp pertamina hulu agar menunda pengurangan tenaga kerja

Tabel 6. Hasil Normalisasi

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

D1	D2	D3
Rapat paripurrna agenda pelantikan pergantian wakil ketua dprd	pemerintah kota pemko pekanbaru mengumpulkan data masyarakat miskin positif covid	himbauan badan operasi bumi siak pusako bsp pertamina hulu menunda pengurangan tenaga kerja

Tabel 7. Hasil Remove Stopword

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

D1	D2	D3
Rapat paripurrna agenda lantik ganti wakil ketua dprd	perintah kota pemko pekanbaru kumpul data masyarakat miskin positif covid	himbauan badan operasi bumi siak pusako bsp pertamina hulu tunda kurang tenaga kerja

Tabel 8. Tabel Hasil Stemming

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

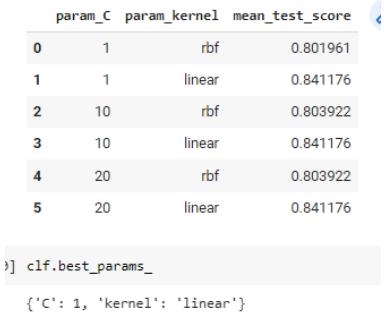
Term	TF			DF	IDF	TF-IDF		
	D1	D2	D3			D1	D2	D3
agenda	1	0	0	1	ln(4/2)+1 = 1.693	1.693	0	0
badan	0	0	1	1	ln(4/2)+1 = 1.693	0	0	1.693
bsp	0	0	1	1	ln(4/2)+1 = 1.693	0	0	1.693
bumi	0	0	1	1	ln(4/2)+1 = 1.693	0	0	1.693
covid	0	1	0	1	ln(4/2)+1 = 1.693	0	1.693	0
data	0	1	0	1	ln(4/2)+1 = 1.693	0	1.693	0
dprd	1	0	0	1	ln(4/2)+1 = 1.693	1.693	0	0
ganti	1	0	0	1	ln(4/2)+1 = 1.693	1.693	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
kerja	0	0	1	1	ln(4/2)+1 = 1.693	0	0	1.693
ketua	1	0	0	1	ln(4/2)+1 = 1.693	1.693	0	0
kota	0	1	0	1	ln(4/2)+1 = 1.693	0	1.693	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
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
operasi	0	0	1	1	ln(4/2)+1 = 1.693	0	0	1.693
paripurrna	1	0	0	1	ln(4/2)+1 = 1.693	1.693	0	0
pekanbaru	0	1	0	1	ln(4/2)+1 = 1.693	0	1.693	0
pemko	0	1	0	1	ln(4/2)+1 = 1.693	0	1.693	0
perintah	0	1	0	1	ln(4/2)+1 = 1.693	0	1.693	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
rapat	1	0	0	1	ln(4/2)+1 = 1.693	1.693	0	0
siak	0	0	1	1	ln(4/2)+1 = 1.693	0	0	1.693
tenaga	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	1	0	0	1	ln(4/2)+1 = 1.693	1.693	0	0

Tabel 9. Tabel hasil TF-IDF

4.3. Klasifikasi *Support Vector Machine*

Pada tahap ini melibatkan pemilihan fungsi kernel dan parameter C. Ada banyak parameter yang digunakan dalam model SVM ini, namun parameter yang memberkan pengaruh paling besar dalam hasil klasifikasi adalah kernel dan parameter C. Untuk mendapatkan nilai paramter C dan kernel yang terbaik dapat dilakukan menggunakan metod *Grid Search* dan *Cross Validation.* Metode *grid search* ini bertujuan untuk menentukan nila C dan Kernel secara manual dengan melakukan pencarian grid secara berulang ulang. Hasil pencarian parameter C dan Kernel terbaik didapatkan yaitu 1 dengan kernel terbaik yakni kernel Linear.



	param_C	param_kernel	mean_test_score
0	1	rbf	0.801961
1	1	linear	0.841176
2	10	rbf	0.803922
3	10	linear	0.841176
4	20	rbf	0.803922
5	20	linear	0.841176

```
)] clf.best_params_
{'C': 1, 'kernel': 'linear'}
```

Gambar 3. Pencarian Parameter Terbaik

Berikut hasil dari klasifikasi dengan skenario pembagian data yang berbeda beda menggunakan Parameter C=1 dan Kernel Linear.

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.

Predicted Classification	*Actual Classification*		
	Demokrasi Positive	Kemiskinan Positive	Ketenagakerjaan Positive
Demokrasi Negative	45	3	6
Kemiskinan Negative	1	41	5
Ketenagakerjaan Negative	5	2	45

Tabel 10. Hasil *Confusion Matrix* 70% data latih dan 30% data uji

Perhitungan akurasi $= \frac{45+41+45}{45+3+6+1+41+5+5+2+45} \; x \; 100\% = \frac{131}{153} \; x \; 100\%$
$= 86\%$

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.

Predicted Classification	*Actual Classification*		
	Demokrasi Positive	Kemiskinan Positive	Ketenagakerjaan Positive
Demokrasi Negative	28	1	6
Kemiskinan Negative	0	30	2
Ketenagakerjaan Negative	6	1	28

Tabel 11. Hasil *Confusion Matrix* 80% data latih dan 20% data uji

Perhitungan akurasi $= \frac{28+30+28}{28+1+6+0+30+2+6+1+28} \; x \; 100\% = \frac{86}{102} \; x \; 100\%$
$= 84\%$

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.

Predicted Classification	*Actual Classification*		
	Demokrasi Positive	Kemiskinan Positive	Ketenagakerjaan Positive
Demokrasi Negative	14	2	0
Kemiskinan Negative	0	16	0
Ketenagakerjaan Negative	2	2	15

Tabel 12. Hasil *Confusion Matrix* 90% data latih dan 10% data uji

Perhitungan akurasi $= \frac{14+16+15}{14+2+0+0+16+0+2+2+15} \; x \; 100\% = \frac{45}{51} \; x \; 100\%$
$= 88\%$

5. Kesimpulan

Berdasarkan hasil dari penelitian yang telah dilakukan, dapat disimpulkan beberapa hal sebagai berikut:

1. Metode *Support Vector Machine* terbukti dapat digunakan dalam proses klasifikasi berita.
2. Nilai parameter C terbaik berada di angka 1 dan kernel terbaik yaitu kernel Linear.
3. Hasil akurasi tertinggi yang didapatkan pada skenario pembagian data 90% dan 10% yaitu sebesar 88% dengan data yang digunakan sebanyak 510 data berita.

Daftar Pustaka

[1] "Kemkominfo: Pengguna Internet di Indonesia Capai 82 Juta," 2014. .
[2] A. Tria and B. Achmad, "Sistem Koreksi Kata Dan Pengenalan Struktur Kalimat Berbahasa Indonesia Dengan Pendekatan Kamus Berbasis Levenshtein Distance," *J. SPIRIT*, vol. 9, no. 1, pp. 48–61, 2017.
[3] K. R. Prilianti and H. Wijaya, "Aplikasi text mining untuk automasi penentuan tren topik skripsi dengan metode K-Means Clustering," *J. Cybermatika*, vol. 2, no. 1, 2014.
[4] D. Ariadi and K. Fithriasari, "Klasifikasi Berita Indonesia Menggunakan Metode Naive Bayesian Classification dan Support Vector Machine dengan Confix Stripping Stemmer," *J. SAINS DAN SENI ITS Vol. 4, No.2*, vol. 4, no. 2, pp. 248–253, 2015.
[5] A. Tumanggor and P. S. Hasugian, "Penerapan Data Mining Untuk Memprediksi Tingkat Kemampuan Anak Dalam Mengikuti Mata Pelajaran Dengan Metode C4. 5 Pada SDN 105351 Bakaran Batu," *J. Nas. Komputasi dan Teknol. Inf.*, vol. 4, no. 1, pp. 57–63, 2021.
[6] A. Manur, "IMPLEMENTASI ALGORITMA MANBER PADA PERSAMAAN MAKNA BAHASA INDONESIA DAN MELAYU BERBASIS ANDROID," UNIVERSITAS SUMATERA UTARA, 2017.
[7] S. N. Asiyah and K. Fithriasari, "Klasifikasi Berita Online Menggunakan Metode Support Vector Machine Dan K-Nearest Neighbor Online News Classification Using Support Vector Machine and K-Nearest," *J. Sains dan Seni ITS*, vol. 5, no. 2, 2016.
[8] K. D. Y. Sari, "KATEGORISASI TEKS DENGAN METODE KLASIFIKASI SUPPORT VECTOR MACHINE (SVM) (TEXT CATEGORIZATION WITH SUPPORT VECTOR MACHINE Kategorisasi teks adalah suatu proses pengklasifikasian dokumen-dokumen ke dalam satu atau lebih kategori yang telah didefinisikan," 2006.
[9] A. S. Nugroho, A. B. Witarto, and D. Handoko, "Support Vector Machine in Bioinformatics," *Proceeding Indones. Sci. Meeiting Cent*, 2003.
[10] N. Mz, "Klasifikasi Tweet Berdasarkan Topik Berita Dengan Metode Support Vector Machine (Svm)," Universitas Islam Negeri Sultan Syarif Kasim Riau, 2021.
[11] F. Rahman, "KLASIFIKASI EMOSI UNTUK TEKS BERBAHASA INDONESIA PADA PENGGUNA TWITTER MENGENAI PRESIDEN JOKO WIDODO," Institut Teknologi Sepuluh Nopember, 2018.