Analisis Dan Implementasi *Support Vector Machine* Dengan *String Kernel* Dalam Melakukan Klasifikasi Berita Berbahasa Indonesia

Analysis and Implementation Support Vector Machine With String Kernel for Classification indonesian news

Honakan [1], Adiwijaya [2], Said Al Faraby [3]
[1, 2, 3] Program Studi S1 Ilmu Komputasi, Fakultas Informatika, Universitas Telkom
Jl. Telekomunikasi No. 1, Ters. Buah Batu Bandung 40257 Indonesia
[1] honakangian@gmail.com, [2] adiwijaya@telkomuniversity.ac.id, [3] saidalfaraby@telkomuniversity.ac.id

Abstrak - Kebutuhan analisis *text mining* sangat diperlukan dalam menangani teks yang tidak terstruktur tersebut. Salah satu kegiatan penting dalam text mining adalah klasifikasi atau kategorisasi teks. Analisis text mining ini dilakukan agar mempermudah kita dalam mengambil informasi atau mengelolah informasi yang begitu banyak dari dunia internet atau digital, salah satu nya dengan melakukan klasifikasi dengan data yang sudah tersedia. Kategorisasi teks memiliki berbagai cara untuk melakukan pendekatan antara lain pendekatan *probabilistic*, *support vector machine*, *artificial neural network*, atau *decision tree classification*. Dalam pembelajaran statistik. Support Vector Machine dipilih karena metode ini memiliki kelebihan dalam bidang klasifikasi dengan bantuan kernel. Pada tugas akhir ini support vector machine akan mengelompokkan berita berdasarkan topik menjadi 3 bagian atau *class* yaitu : pemerintahan, ekonomi dan olahraga. Kernel pada Support Vector Mechine akan di kombinasikan dengan *stopword*, *tokenisasi*, *tf-idf*, *chi-square* diharapkan memudahkan untuk mengenali berita tersebut tergelong masuk ke dalam kelas topik yang seharusnya.
Dengan trik kernel dan bantuan metode pembobotan, Dokumen Frekuensi,*Chi square* diharapkan dapat membantu klasifikasi teks dengan baik yang non linear serta mampu meningkatkan akurasi, dengan demikian klasifikasi dengan metode support vektor machine dapat akurasi tertinggi dengan kombinasi *stopword*, *tokenizing*, *term frequency* & *chi-square* 47,43 %.
Kata Kunci : : *text mining, support vector machine, tf-idf, chi square, stopword, tokenisasi.*

Abstract *- The need for text mining analysis is needed in handling the unstructured text. One of the important activities in text mining is the classification or categorization of texts. Text mining analysis is done in order to facilitate us in taking information or managing so much information from the world of the Internet or digital, one of them by classifying the data already available. Text categorization has various ways to approach, among others, probabilistic approach, support vector machine, artificial neural network, or decision tree classification. In statistical learning. Support Vector Machine is chosen because this method has advantages in the field of classification with the help of the kernel. In this final project support vector machine will group news based on topic into 3 part or class that is: government, economy and sport. The kernel of Support Vector Mechine will be combined with stopword, tokenisasi, tf-idf, chi-square is expected to make it easier to recognize the news rolled into the class of topics that should be.*
With the kernel trick and the help of weighting method, Document Frequency, Chi square is expected to help classify the text with both non-linear and can improve accuracy, thus classification with support vector machine method can be the highest accuracy with the combination of stopword, tokenizing, term frequency & chi- square 47.43%
Keyword *: text mining, support vector machine, tf-idf, chi square, stopword, tokenisasi.*

I. PENDAHULUAN

Berita adalah informasi yang sering diakses orang melalui media digital baik itu surat kabar, televisi, maupun di dalam dunia internet. Kebutuhan akan informasi teks berita dikarenakan banyaknya masyarakat sekarang yang ingin mengikuti perkembangan informasi yang sedang terjadi dunia perpolitikan, olahraga, dan entertaiment. Dengan permintaan yang cukup tinggi ini pun dimanfaatkan oleh para media, sehingga di pasaran membuat banyak penyedia atau media yang menghasilkan infromasi teks dalam bentuk teks berita. Salah satu yang paling diminati saat ini adalah berita, orang akan lebih sering mengakses berita terutama dalam bidang olahraga, hal ini disebabkan karena masyarakat ingin mengetauhi update tentang olahraga di bidang sepakbola, basket, tenis, dll,. Berita olahraga adalah berita yang paling sering meng-update atau paling sering menghasilkan data atau informasi berupa text, hal ini disebabkan karena banyaknya pertandingan yang berlangsung atau yang diadakan dalam hitungan hari maupun minggu. Banyaknya berita yang disajikan dalam bentuk teks maupun video mengakibatkan data yang dihasilkan begitu banyak dan tentunya tidak terstruktur. Data tidak terstruktur adalah data yang tidak mudah diklasifikasi dan dimasukan kedalam sebuah kotak dengan rapi, Contohnya adalah foto, gambar grafis, streaming instrument data, webpages, pdf, Power Point presentations, konten blog dan lain sebagainya, sehingga dibutuhkan

suatu *Text Mining* untuk menangani data yang tidak terstruktur. Text mining membantu untuk mengelolah data tersebut yang tidak tertruktur untuk dikelolah atau di proses dengan mencoba menenemukan pola pola yang dapat digali dan digunakan untuk mencari informasi dari data yang tidak terstruktur tersebut. Ada beberapa teknik untuk melakukan text mining yaitu naive bayes, *single pass clustering*, *support vecktor machine* dll. Pada tugas Akhir ini kita akan menggunakan metode SVM. Konsep SVM dapat dijelaskan secara sederhana sebagai usaha mencari hyperplane terbaik yang berfungsi sebagai pemisah dua buah class pada input *space*. Metode ini berakar dari teori pembelajaran statistik yang hasilnya sangat menjanjikan untuk memberikan hasil yang lebih baik daripada metode yang lain [2].

II. LANDASAN TEORI

A. Data mining.

Data *mining* sebagai proses untuk mendapatkan informasi yang berguna dari gudang basis data yang besar. Data mining juga dapat diartikan sebagai pengekstrakan informasi baru yang diambil dari sekumpulan bongkahan data yang besar yang membantu pengambilan keputusan. Istilah data mining di sebut juga knowledge discovery (KDD) [2]. Secara sederhana bisa dikatakan data mining adalah teknik untuk mengambil dan mengelolah data yang begitu besar untuk mendapatkan informasi. Dalam klasifikasi Salah satu teknik yang dibuat dalam data mining adalah bagaimana menelusuri data yang ada untuk membangun sebuah model, kemudian menggunakan model tersebut agar dapat mengenali pola data yang lain yang tidak berada dalam basis data yang tersimpan. Kebutuhan untuk prediksi juga dapat memanfaatkan teknik ini. Dalam data mining pengelempokkan data juga bisa di lakukan. Tujuannya adalah agar kita dapat mengetahui pola universal data-data yang ada. Anomali data transaksi juga perlu dideteksi untuk dapat mengetahui tindak lanjut berikutnya yang dapat diambil [2]. Pekerjaan yang berkaitan dengan data mining dapat dibagi menjadi empat kelompok yaitu model prediksi (prediction modelling), analisis kelompok (cluster analysis), analisis asosiasi (association analysis) dan deteksi anomali (anomali detection).

B. Machine learning

Machine learning adalah salah satu ilmu pengetahuan yang mencakup tentang pengembangan dan perancangan algoritma yang bertujuan agar memungkinkan sebuah program komputer mampu mengembangkan perilaku yang didasarkan pada data empiris [10]. Bisa dikatakan machine learning adalah ilmu yang bertujuan untuk membuat sautu mesin mampu melakukan pembelajaran atau pekerjaannya sendiri dari pembangunan model yang dibangun dari kumpulan data. Sehingga data sangat diperlukan apabila ingin membangun sebuah model, model yang telah dibangun tersebut akan digunakan sebagai rule pada mesin, sehingga mesin mampu untuk melakukan pembelajaran dan pekerjaan nya sendiri.
Bisa dikatakan pembelajaran mesin adalah cabang ilmu dari kecerdasan buatan yang bertujuan agar suatu mesin atau komputer dapat melakukan pekerjaannya sendiri dengan memanfaatkan data yang sudah ada, dengan memanfaatkan data maka bisa dibangun sebuah rule atau algoritma agar mesin dapat mengambil keputusan sendiri dari rule atau algoritma yang telah dibangun tersebut. Ada beberapa cara yang biasa digunakan untuk membangun sebuah kaidah (rule) atau algoritma pada mesin, diantaranya adalah dengan memanfaatkan ilmu statistika, untuk melakukan pendekatannya bisa menggunakan pendekatan secara fisiologi yaitu seperti meniru jaringan syaraf tiruan atau biasa yang di sebut ANN (Artifical Neural Network)

C. Preprecosssing

Set data yang akan di proses dengan metode metode dalam data mining sering kali harus melalui pekerjaan awal, ini adalah tahapan awal ketika kita ingin mengelolah data input-an yang secara keseluruhan terpisah dari metode dalam data mining. Masalah masalah seperti jumlah suatu populasi data yang terlalu besar, banyaknya data yang menyimpang (anomali data), dimensi yang terlalu tinggi, banyaknya fitur yang tidak berkontribusi besar dan lain lain menjadi pemicu munculnya pemrosesan awal (pre-processing) yang harus diterapkan padaa set data sebelum akhirnya dilepas untuk di proses di dalam data mining. Beberapa pekerjaan yang umum dilakukan sebagai awal preprocessing pada set data adalah agregasi, penyampelan, pengurangan dimensi, pemilihan fitur, diskretisasi dan binerisasi, dan transformasi variabel [2] Jadi bisa di sumpulkan Preprocessing adalah teknik maupun strategi yang bertujuan untuk membuat suatu data lebih mudah untuk dikelolah atau cocok untuk digunakan pada data mining yang tentunya bertujuan agar meningkatkan hasil dari analisis data mining.

D. Pembobotan

Pembobotan dilakukan agar membuat proses pelatihan dan penggunaan fungsi klasifikasi lebih efisien. dengan mengurangi jumlah kata yang digunakan. Hal ini juga dapat meningkatkan akurasi hasil klasifikasi. Pada klasifikasi ini penulis menggunakan fitur pembobotan. Pemilihan fitur memiliki 2 tujuan utama. Pertama, membuat proses pelatihan dan penggunaan fungsi klasifikasi lebih efisien dengan mengurangi jumlah kata yang digunakan. Kedua, hal ini dapat meningkatkan akurasi hasil klasifikasi. Pada tugas akhir ini penulis menggunakan metode pembobotan sebagai fitur untuk mengubah data menjadi nominal

1. **Fitur *term frequency (tf)***
Akan memperhitungkan berapa banyak kemunculan sebuah fitur dalam suatu dokumen

Dokumen	Fitur (kemunculan)
Dokumen 1	Pajak (3) cukai (4) uang (6) menteri (2)
Dokumen 2	Pilkada (3) partai (4) uang (2)
Dokumen 3	Gol (4) pertandingan (2) pemimpim (3)

2. **Fitur *invers dokument frequency (idf)***
Fitur akan dihitung berdasarkan kemunculan fitur pada sebuah dokumen dibagi dengan jumlah dokumen yang memiliki fitur tersebut.

Rumus : $\text{idf} = \frac{1}{df}$ atau $\text{idf} = \log\left(\frac{N}{df}\right)$ (1)

Dokumen	Fitur (kemunculan)
Dokumen 1	Pajak $(1/4)$ cukai $(1/4)$ uang $(1/4)$ menteri $(1/4)$
Dokumen 2	Pilkada $(1/3)$ partai $(1/3)$ uang $(1/3)$
Dokumen 3	Gol $(1/3)$ pertandingan $(1/3)$ pemimpim $(1/3)$

3. Chi Square (Chi)
Chi Square adalah salah satu metode pembobotan dan juga sebagai uji komparatif non parametris yang dapat dilakukan pada dua variabel, untuk menguji ada atau tidak adanya keterikatan antara variabel yang satu dengan variabel yang lain, dengan memanfaatkan nilai harapan dan kemunculan atribut dalam setiap dokumen untuk mendapatkan nilai chi-square dari setiap atribut.

$$x_p^2 \sum_{tf} \frac{(f_{ij} - E_{ij})^2}{E_{ij}} \quad (2)$$

Keterangan
f_O = frekuensi awal hasi observasi
f_e = nilai harapan

$$\text{Rumus } f_e : = \frac{jumlah\ baris \times jumlah\ kolom}{total\ keseluruhan} \quad (3)$$

Dokumen	Atribut								Total	
	pajak		cukai		uang		menteri			
	f_O	f_e	f_O	f_e	f_O	f_e	f_O	f_e	f_O	f_e
Dokumen 1	3	4,16	4	2,91	6	4,16	2	3,75	15	15
Dokumen 2	2	2,5	1	1,75	3	2,5	3	2,25	9	9
Dokumen 3	5	3,3	2	2,3	1	3,3	4	3	12	12
Total	10		7		10		9		36	36

E. Support Vector Machine

Konsep Support Vector Machine dapat dijelaskan secara sederhana sebagai usaha mencari hyperplane terbaik yang berfungsi sebagai pemisah dua buah class pada input space. Gambar 2.3 memperlihatkan beberapa pattern yang merupakan anggota dari dua buah class : positif (dinotasikan dengan +1) dan negatif (dinotasikan dengan –1). Pattern yang tergabung pada class negatif disimbolkan dengan kotak, sedangkan pattern pada class positif, disimbolkan dengan lingkaran. Proses pembelajaran dalam problem klasifikasi diterjemahkan sebagai upaya menemukan garis (hyperplane) yang memisahkan antara kedua kelompok tersebut. Berbagai alternatif garis pemisah (discrimination boundaries) ditunjukkan pada Gambar 1a [12]. Hyperplane pemisah terbaik antara kedua class dapat ditemukan dengan mengukur margin hyperplane, dan mencari titik maksimalnya. Margin adalah jarak antara hyperplane tersebut dengan data terdekat dari masing-masing class. Subset data training set yang paling dekat ini disebut sebagai support vector. Garis solid pada Gambar 2.4 menunjukkan hyperplane yang terbaik, yaitu yang terletak tepat pada tengah-tengah kedua class, sedangkan titik kotak dan lingkaran yang berada dalam lingkaran hitam adalah support vector. Upaya mencari lokasi hyperplane optimal ini
Usaha yang digunakan untuk menemukan atau mencari letak atau lokasi hyperplane ini merupakan inti dari proses pada *Support Vector Machine*. Jarak diantara 2 *hyperplane* adalah $\frac{2}{\|w\|}$, jadi untuk memaksimumkan jarak antar kelas maka nilai $\|w\|$ harus diminimumkan. SVM memaksimumkan margin disekitar garis pemisah *hyperplane*. Secara matematika, formulasi problem optimisasi SVM untuk kasus klasifikasi linier adalah :

$$\min \frac{1}{2}\|w\|^2$$

Dimana

$$y_i\ (wx_i + b) \geq 1, i = 1,..,l, \quad (4)$$

&

$$w.x + b\ 1 \leq -1 \quad (5)$$

Dimana xi merupakan data masukan, i merupakan jumlah data dan yi merupakan keluaran, sedangkan w dan b merupakan parameter yang akan dicari nilainya[\

F. Support Vector Machine Non Linier

Pada awalnya machine learning dihadirkan atau diciptakan sebagai pemisah dua kelas dengan memanfaatkan garis pemisah (linear) atau yang dikenal dengan *hyperplane*, namun semakin perkembangan jaman masalah yang dihadapi juga semakin kompleks, semakin banyak masalah dengan non linear, sehingga membutuhkan bantuan string kernel atau fungsi kernel untuk mengatasi permasalahan yang non linier tersebut.
Kernel trick adalah salah satu cara yang dapat membantu untuk memudahkan dalam melakukan klasfikasi dalam bentuk non linear, cukup hanya memahami atau mengetahui fungs kernel apa yang cocok digunakan dalam kasus non linear yang kita hadapi. Kernel digunakan untuk memetakan inputan di *input space* ke dalam *feature space* membuat pemisah antar kelas bukan lagi garis lurus merupakan garis lengkung atau bidang dalam dimensi yang lebih tinggi.
Tentu pendekatan dengan kernel ini berbeda dengan yang metode lain karena biasanya pendekatan dengan metode lain mengurangi dimensi awal untuk menyederhanakan proses komputasi.
.Menurut Hsu, dkk (2010) berikut ini adalah beberapa fungsi kernel yang umum digunakan yaitu:

a. Kernel linier :
$$K(xi,x) = xi^T x \quad (6)$$
b. *Polynomial* :
$$K(xi,x) = (\gamma.xi^T x+r)p\ ,\gamma > 0 \quad (7)$$
c. *Radial Basis Function* :
$$K(xi,x)=\exp(-\gamma \mid xi-x \mid^2),\gamma>0 \quad (8)$$
d. *Sigmoid kernel* :
$$K(xi,x)= \tanh(\gamma xi^T+r) \quad (9)$$

kernel linier digunakan ketika data yang akan diklasifikasi dapat dipisahkan dengan sebuah garis/*hyperplane* sedangkan kernel non-linier digunakan ketika data hanya dapat dipisahkan dengan garis lengkung atau sebuah bidang pada ruang dimensi tinggi.

G. Performansi Sistem

Pengukuran performansi klasifikasi akan menggunakan matrix confusion dan akurasi sebagai parameter yang menjadi patokan. Berikut adalah rumus akurasi dan matrix confusion digunakan :

	Predicted class = YES	Predicted class = NO
Actual class = YES	True positive	False negative
Actual class = No	False positive	True negative

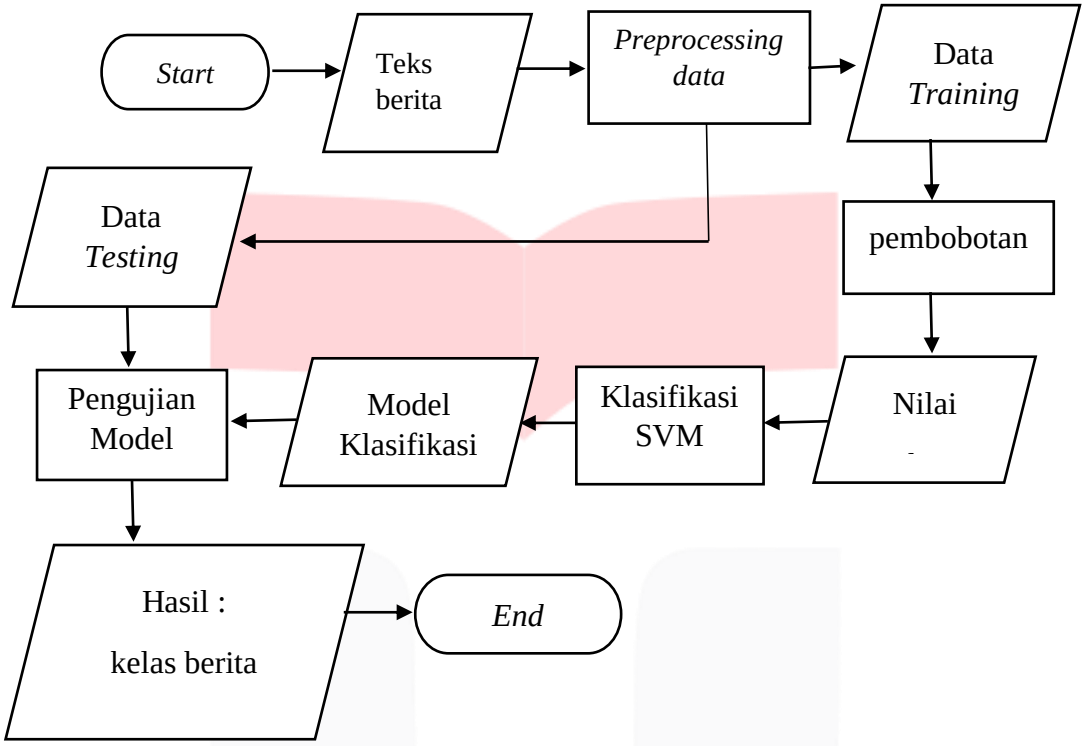
True positive : Jumlah dokumen yang diklasifikasikan dengan benar
False positive : Jumlah dokumen uji yang seharusnya masuk dalam kategori
True negative : Jumlah dokumen uji yang dengan benar tidak dimasukkan dalam kategori
False negative : Jumlah dokumen uji yang seharusnya masuk dalam kategori tetapi tidak dimasukkan dalam kategori

$$\text{Akurasi} = \frac{\text{Jumlah semua status deteksi yang benar}}{\text{Jumlah total data referensi}} \quad (10)$$

Presisi dapat diartikan sebagai nilai keandalan dari data sistem yang diperoleh. Sedangkan *recall* adalah nilai pencapaian atau keberhasilan dalam menemukan kembali suatu informasi. Akurasi digunakan untuk mengukur tingkat kesesuaian semua data yang benar dengan total data referensi.

III. PERANCANGAN SISTEM

Pada Tugas akhir ini penulis akan menggunakan metode Support Vector Machine dengan string kernel untuk mengklasifikasi data data berita. Berikut merupakan gambaran umum dengan flowchart yang akan dilakukan dalam penilitian kali ini



Gambar 1 flowchart sistem secara umum

Mengacu pada sistem penelitian **Error! Reference source not found.**, detail rancangan sistem dijelaskan pada beberapa tahap sebagai berikut:

- Pada tahap pertama melakukan *preprocessing* data yaitu mebersihkan dataset agar sistem lebih mudah dalam melakukan komputerisasi data teks dan menghilangkan kata yang kurang penting yaitu seperti kata keterangan pada teks berita.
  Tahap Tokenizing adalah step preprocessing yang betujuan agar tahap pemotongan string input berdasarkan tiap kata yang menyusunnya. Contohnya
  Teks input : managemen pengatahuan adalah sebuah konsep baru di dunia bisnis
  Hasil : - managemen - adalah - konsep
  - pengetahuan - sebuah - baru
  Tahap stopword Pada tahap ini untuk bertujuan untuk mempersentasikan suatu dokumen dapat mendeskripsikan isi dari suatu dokumen untuk membedakan isi dokumen lain, dalam suatu istilah (*term*) akan mencari jumlah dokumen yang dianggap paling relevan didalam suatu inputan (*query*), suatu *term* yang sering ditampilkan atau digunakan diangap sebagai *stopword.* Contoh: Operator Logika *or, not, a nd* dan sebaginya.
- Tahap kedua yaitu membagi dataset menjadi data training dan data testing.
- Tahap ketiga yaitu melakukan melakukan metode pembobotan pada setiap kata atau term di dalam dokumen
- Tahap keempat yaitu membangun model klasifier dari nilai pembobotan

- Tahap kelima yaitu melakukan klasifikasi pada data testing dengan menggunakan model klasifier yang telah dibangun.
- Tahap terakhir yaitu mengevaluasi sistem dengan menghitung nilai akurasi.

IV. PENGUJIAN DAN ANALISIS

Pada bab ini memuat pembahasan tentang hasil pengujian yang telah dilakukan oleh penulis berdasarkan skenario pengujian yang ada tertera pada bab sebelumnya, pengujian sistem dilakukan dari dataset yang telah di preprocessing hingga diklasifikasi ke dalam support vector mechine, dan akan berisikan analisis dari penulis terhadap hasil klasfikasi yang sebelumnya akan diberikan metode berbeda dalam melakukan pembobotan, Analisis pada akhir pengujian dilakukan agar penulis bisa mengambil kesimpulan dari beberapa metode term, dan mencari metode terbaik dalam melakukan klasifikasi teks dokumen berbahasa Indonesia

A. Pengujian Skenario Menggunakan fungsi kernel linier

pengujian ini dilakukan dengan penjelasan skenario diatas dengan beberapa kombinasi metode pembobotan dengan ekstraksi fitur. Pada percobaan ini pengamat mengunakan fungsi kernel linear untuk pengujian 3 skenario

Tabel 1 Hasil skenario menggunakan fungsi linier dengan 300 data latih

	Ekstraksi fitur dan metode pembobotan					
Skenario	*stopword*	*tokenizing*	*TF*	*idf*	*Chi square*	Akurasi (%)
Skenario 1	✓	✓	✓			23,6
Skenario 2	✓	✓	✓	✓		37,6431
Skenario 3	✓	✓	✓		✓	38,774

Tabel 2 Hasil skenario menggunakan fungsi linier dengan 600 data latih

	Ekstraksi fitur dan metode pembobotan					
Skenario	*stopword*	*tokenizing*	*TF*	*idf*	*Chi square*	Akurasi (%)
Skenario 1	✓	✓	✓			39,6
Skenario 2	✓	✓	✓	✓		41,33
Skenario 3	✓	✓	✓		✓	44,4

B. Pengujian Skenario Menggunakan fungsi kernel Radial Basis Function

Pengujian ini dilakukan dengan penjelasan skenario diatas dengan beberapa kobinasi metode pembobotan dengan ekstraksi fitur. Pada percobaan ini pengamat mengunakan fungsi kernel linear untuk pengujian 5 skenario

Tabel *3* Hasil skenario menggunakan fungsi kernel *Radial Basis Function dengan 300 data*

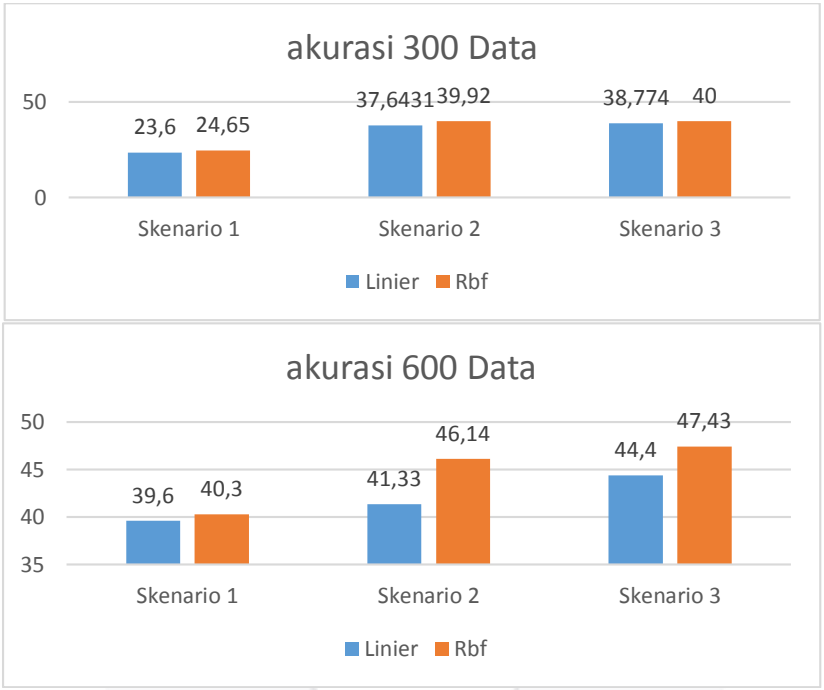
	Ekstraksi fitur dan metode pembobotan					
Skenario	*stopword*	*tokenizing*	*TF*	*idf*	*Chi square*	*Akurasi (%)*
Skenario 1	✓	✓	✓			*24,65*
Skenario 2	✓	✓	✓	✓		*39,92*
Skenario 3	✓	✓	✓		✓	*40*

Tabel *4* Hasil skenario menggunakan fungsi kernel *Radial Basis Function dengan 600 data*

	Ekstraksi fitur dan metode pembobotan					
Skenario	*stopword*	*tokenizing*	*TF*	*idf*	*Chi square*	Akurasi(%)
Skenario 1	✓	✓	✓			40,30
Skenario 2	✓	✓	✓	✓		46,14
Skenario 3	✓	✓	✓		✓	47,43

C. Analisis

Beberapa skenario percobaan telah dilakukan dan didapatkan hasil yang berbeda beda oleh karena di kombinasi yang dilakukan pada setiap percobaan juga berbeda. Stoprword dalam setiap skenario dapat membantu mempercepat dan membuang kata yang tidak bersifat unik. Tokenisasi berperan dalam memisahkan kata per kata dari kalimat yang terdapat pada berita. Term frequensi untuk menghitung jumlah kemunculan atribut kata unik di dalam suatu dokumen, term frequensi sangat dibutuhkan sebagai nilai awal dalam melakukan pembobotan untuk tf-idf dan chi-square. Dari perbandingan skenario 2 & 3 diketahui hasil akurasi dengan pembobotan chi square lebih unggul. Pengaruh data latih juga sangat mempengaruhi jumlah akurasi, semakin banyak data latih maka kemungkinan semakin banyak kata unik yang didapatkan sehingga dapat mengenali atau mengelompokkan data testing dengan mudah, penyebab rendahnya akurasi disebabkan terlalu besarnya cakupan dalam satu topik berita olahraga contohnya data unik untuk cabang olahraga catur akan sangat berbeda dengan cabang olahraga sepakbola maupun sebaliknya, sehingga dibutuhkan data latih yang besar sebagai database kata unik untuk semua cakupan.



Gambar *2* Hasil akurasi semua skenario

Hasil dari beberapa skenario 1,2 & 3 dapat dilihat perbedaan akurasi dalam perbedaan kombinasi motede dan eksraksi ciri. Dari beberapa kombinasi didapatkan bahwa metode pembobotan dengan metode chi square mendapatkan hasil akurasi tertinggi dibandingkan dengan kombinasi metode pembobotan yang lainnya, data juga berpengaruh dalam meningkatkan akurasi hal itu dilihat dari perbedaaan jumlah pemberian data latih pada beberapa percobaan itu, disebabkan kurangnya data latih sehingga kata yang besifat unik tidak dapat mencakup seluruh cakupan didalam topik kelas berita tersebut. Kata unik yang sedikit akan menyebabkan support vecktor mechine sulit melakukan proses klasifikasi dengan baik, sehingga kemungkinan salah untuk mengelompokkan berita lebih besar dibandingkan dengan pemberian data latih yang lebih banyak. Preprocessing pada setiap skenario juga sangat dibutuhkan karena masalah seperti jumlah suatu populasi data yang terlalu besar, banyaknya data yang menyimpang (anomali data), dimensi yang terlalu tinggi, banyaknya fitur yang tidak berkontribusi besar dan lain lain menjadi pemicu munculnya pemrosesan awal (pre-processing) yang harus diterapkan padaa set data sebelum akhirnya dilepas untuk di proses di dalam data mining sehinggan dilakukan stopword dan tokenisasi. Pembobotan dilakukan sebagai salah cara untuk mengubah diskrit menjadi data numerik, untuk memberikan nilai dari setiap kata yang ada pada dokumen tersebut. Tf-idf dan chi square adalah metode pembobotan yang diberikan pada skenario tersebut, dan dapat dilihat chi-square lebih baik dalam hal akurasi, dikarenakan salah satu jenis uji komparatif non parametris yang dilakukan pada dua variabel, di mana skala data kedua variabel adalah nominal.

V. KESIMPULAN DAN SARAN

Setelah dilakukan percobaan dari beberapa skenario yang telah dilakukan oleh penulis, maka dapat ditarik kesimpulan yang mampu menggambarkan percobaan ini, yaitu preprocessing sangat diperlukan dalam melakukan klasifikasi dalam bentuk teks, karena membantu untuk mengurangi fitur yang terlalu banyak, sehingga bisa membantu proses klasifikasi agar lebih cepat dalam hal waktu, dan juga preprocessing membantu. Pembobotan tf-idf dan tf-chi square sangat membantu dalam melakukan klasifikasi dengan memberi nilai bobot pada kata unik yang ada pada berita, sebagai model atau kata kunci untuk menentukan class pada tahap pengujian data dan dapat meningkatkan akurasi. Kombinasi dari beberapa fitur akan mempengaruhi akurasi walaupun tidak begitu signifikan. Kernel "RBF" pada support vector mechine mampu menambah hasilakurasi dikarenakan data yang digunakan *multiclass*

Adapun saran yang penulis ajukan dalam penelitian selanjutnya yaitu Sebaiknya sebelum melakukan proses klasifikasi lakukan *preprocessing* pada data testing. Apabila ingin mendapatkan akurasi yang bagus dalam klasifikasi berita sebaiknya data latih diperbanyak karena cakupan dalam suatu topik juga besar saat mejalankan program klasifikasi sebaiknya menggunakan komputer yang memiliki spek yang bagus agar mempercepat proses kumputasi. Untuk meningkatkan akurasi sebaiknya memuat topik berita lebih spesifik, agar memperkecil ruang lingkup berita tersebut.

REFERENSI

[1] Lodhi, Huma, et al. "Text classification using string kernels." *Journal of Machine Learning Research* 2.Feb (2002): 419-444.

[2] E. prasetyo, "Data Mining konsep dan aplikasi menggunakan Matlab",Yogyakarta: PENERBIT ANDI, 2009

[3] Tanmay Basu and C.A. Murthy "A Feature Selection Method for Improved Document Classification". 2012.

[4] Tang, Jiliang, Salem Alelyani, and Huan Liu. "Feature selection for classification: A review." *Data Classification: Algorithms and Applications*(2014): 37

[5] Dr. rer. nat. Hendri Murfi "MMA10991 Topik Khusus – Machine Learning

[6] prabowo pudjo widodo,Rahmadya trias, Herlaawati "penerapan Data Mining dengan Matlab", Bandung: Rekayasa Sains, 2013.

[7] Thorsten Joachims, "Text Categorization with SupportVector Machines:Learningwith Many Relevant Features

[8] Thorsten Joachims,"transductive inference text clasification using support vector machine

[9] Deng, Shuang, and Hong Peng. "Document classification based on support vector machine using a concept vector model." *2006 IEEE/WIC/ACM International Conference on Web Intelligence (WI 2006 Main Conference Proceedings)(WI'06)*. IEEE, 2006.

[10] Rosyita Ayuning M "SIMULASI DAN ANALISIS KLASIFIKASI GENRE MUSIK BERBASIS FFT DAN SUPPORT VECTOR MACHINE" .2015. Telkom University 6.

[11] Adyatma Bhaskara H "KLASIFIKASI DOKUMEN BERITA MENGGUNAKAN METODE SUPPORT VECTOR MACHINE DENGAN KERNEL RADIAL BASIS FUNCTION" 2014. Institut Pertanian Bogor.

[12] C.A. Murthy, Tanmay Basu, "A Feature Selection Method for Improved Document Classification

[13] Asriyanti Indah Pratiwi, Adiwijaya. 2017. On The Feature Selection and Classification Based on Information Gain for Document Sentiment Analysis, Applied Computational Intelligence and Soft Computing

[14] Mubarok, M.S., Adiwijaya and Aldhi, M.D., 2017. Aspect-based sentiment analysis to review products using Naïve Bayes. In AIP Conference Proceedings (Vol. 1867, No. 1, p. 020060). AIP Publishing.

[15] Aziz, R.A., Mubarok, M.S. and Adiwijaya, A., 2016, September. Klasifikasi Topik pada Lirik Lagu dengan Metode Multinomial Naive Bayes. In Indonesia Symposium on Computing (IndoSC) 2016

[16] Arifin, A.H.R.Z., Mubarok, M.S. and Adiwijaya, A., 2016, September. Learning Struktur Bayesian Networks menggunakan Novel Modified Binary Differential Evolution pada Klasifikasi Data. In Indonesia Symposium on Computing (IndoSC) 2016.

[17] Irene Yulietha, Said Faraby, A. Adiwijaya. 2017. Klasifikasi Sentimen Review Film Menggunakan Algoritma Support Vector Machine. eProceedings of Engineering 4 (3)

[18] Adiwijaya. 2014. Aplikasi Matriks dan Ruang Vektor. Yogyakarta: Graha Ilmu

[19] Adiwijaya, 2016, Matematika Diskrit dan Aplikasinya, Bandung: Alfabeta

[20] Adiwijaya, Salman ANM., Serra, O, Suprijanto, D, Baskoro, ET. 2015. Some graphs in Cf 2 based on f-coloring. Int. J. Pure Appl. Math, 102(2), pp.201-207