Klasifikasi Halaman Web untuk Anak Menggunakan Metode *Weighted Voting Support Vector Machine*

Maulidya Yuniarti*[1], Gita Indah Marthasari[2], Nur Hayatin[3]
[1]Universitas Muhammadiyah Malang, [2,3]Universitas Muhammadiyah Malang
e-mail: maulidyayuniarti@gmail.com*[1]

Abstrak

Saat ini pengguna internet meningkat seiring dengan pesatnya kemajuan teknologi. Banyak diantara mereka merupakan anak usia sekolah yang belum dapat memilah halaman web sesuai dengan usia mereka, untuk itu dibutuhkan peran orang tua untuk mengontrol akses internet pada ana. Tentu saja para orang tua akan kesulitan memilih halaman web yang sesuai dengan usia anak mereka dikarenakan jumlah halaman web yang sangat banyak. Untuk mengatasi permasalahan tersebut dibutuhkan sistem klasifikasi halaman web secara otomatis. Dengan adanya sistem tersebut tentunya juga dapat membuat halaman web yang diakses oleh anak-anak lebih dapat terkontrol sesuai dengan usia mereka. Oleh karena itu pada penelitian ini dilakukan klasifikasi halaman web anak secara otomatis dengan menggunakan algoritma Weighted Voting Support Vector Machine (*WVSVM) yaitu Support Vector Machine (SVM) yang dikombinasikan dengan ekstraksi fitur Latent Semantic Analysis (LSA) dan Web Page Feature Selection (WPFS). Untuk mengetahui performa dari WVSVM, hasil pengujian dari algoritma WVSVM akan dibandingkan dengan hasil pengujian algoritma LSA-SVM dan WPFS-SVM. Dengan menggunakan confussion matrix didapat hasil bahwa WVSVM lebih unggul dibandingkan dua algoritma lainnya yakni mendapatkan hasil 74% untuk akurasi, 76% untuk presisi, 68% untuk recall dan f-measure sebesar 71%, hasil pengujian tersebut merupakan hasil klasifikasi dengan menggunakan kernel RBF dan ratio data 80:20.*

Kata kunci: *klasifikasi halaman web, latent semantic analysis, web page feature selection, svm.*

Abstract

Nowadays internet users are increasing along with the rapid advancement of technology. Many of them are school-age children who have not been able to sort web pages according to their age, for this reason the role of parents is needed to control internet access for you. Of course, parents will have difficulty choosing web pages that are appropriate for their child's age due to the large number of web pages. To overcome these problems, a web page classification system is needed automatically. With this system, of course, it can also make web pages accessed by children more controlled according to their age. Therefore in this study the classification of children's web pages is automatically carried out using the Weighted Voting Support Vector Machine (WVSVM) algorithm, which is Support Vector Machine (SVM) combined with Latent Semantic Analysis (LSA) feature extraction and Web Page Feature Selection (WPFS) . To find out the performance of WVSVM, the test results of the WVSVM algorithm will be compared with the results of testing the LSA-SVM and WPFS-SVM algorithms. By using the confusion matrix, the results show that WVSVM is superior to the other two algorithms, which get 74% for accuracy, 76% for precision, 68% for recall and f-measure by 71%, the test results are the result of classification using the RBF kernel and data ratio of 80:20.

Keywords *web page classification, latent semantic analysis, web page feature selection, svm.*

1. Pendahuluan

Pesatnya perkembangan teknologi komputer saat ini juga diiringi dengan peningkatan pengguna internet dalam kehidupan sehari-hari. Setidaknya, pada tahun 2018 diperkirakan sebanyak 3,6 miliar manusia di seluruh dunia akan menggunakan internet setidaknya sekali dalam satu bulan. Saat ini pengguna internet bukan hanya orang dewasa, tetapi juga anak usia sekolah. Penelitian UK menemukan lebih dari 40% anak usia 5-15 tahun mengakses internet tanpa adanya pengawasan dari Orang Tua. Padahal dengan pengembalian halaman web yang tidak relevan dapat menyebabkan anak-anak tersebut

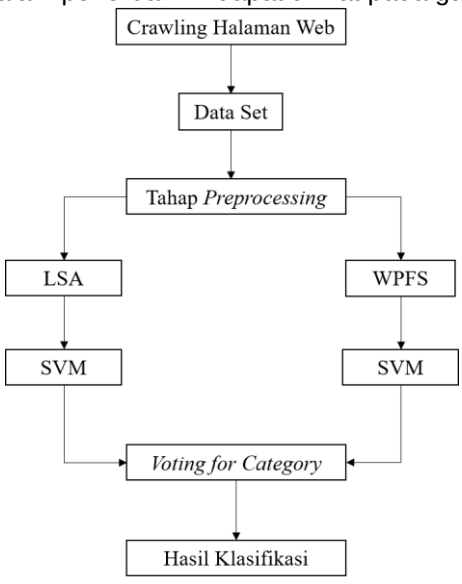
Makalah dikirim 2 Oktober 2019

mengakses halaman web yang tidak sesuai dengan usia mereka. Untuk mengatasi permasalahan tersebut dibutuhkan sistem yang dapat melakukan klasifikasi halaman web secara otomatis. Dengan adanya sistem tersebut tentunya juga dapat membuat halaman web yang diakses oleh anak-anak lebih dapat terkontrol sesuai dengan usia mereka. Beberapa peneliti telah menerapkan SVM dalam klasifikasi halaman web, diantaranya adalah Sun dkk.[1], Dixit dkk.[2], He dkk[3] dan Chen dkk[4]. Chen dkk melakukan klasifikasi halaman web menggunakan metode WVSVM (*Weighted Voting Support Vector Machine*) yang mengkombinasikan SVM dengan penambahan seleksi fitur menggunakan LSA (*Latent Semantic Analysis*) untuk mengekstrak hubungan semantik umum antara kata kunci dengan dokumen, selain menggunakan LSA pada penelitian ini juga menggunakan *Web Page Feature Selection* untuk ektraksi teks fitur dari halaman web. Halaman web yang digunakan sebagai data pada penelitian ini merupakan situs berita olahraga yang diunduh dari halaman web http://udndata.com untuk di klasifikasikan kedalam beberapa kategori yaitu *basketball, baseball, golf, tennis, volleyball, soccer, billiards, football,* dan *Formula 1 Racing.* Setelah melakukan klasifikasi, peneliti melakukan pengujian terhadap sistem dengan menggunakan *F-measure, recall* dan *precision* yang dibandingkan dengan hasil pengujian pada metode LSA-SVM dan BPN. Hasilnya menunjukkan bahwa WVSVM unggul dari metode yang lain yaitu mendapatkan hasil 90% untuk *precision*, 97% untuk *recall* dan 93% untuk nilai *F-measure*.

Dari hasil pengujian pada penelitian yang dilakukan Chen dkk dapat diketahui bahwa metode WVSVM yang mereka usulkan mendapatkan nilai akurasi yang sangat baik. Oleh karena itu, penulis mengusulkan sebuah penelitian dengan menerapkan algoritma *Support Vector Machine* yang dikombinasikan dengan *Latent Semantic Analysis* dan *Web Page Feature Selection* sebagai metode seleksi fitur pada data set yang digunakan oleh Eickhoff dkk[5] yaitu kumpulan halaman web yang didapatkan dari situs http://dmoz-odp.org/. Pada penelitian tersebut Eickhoff dkk melakukan klasifikasi halaman web menjadi dua kelas target yaitu halaman web untuk anak dan untuk umum menggunakan *logistic regression* sebagai algoritma untuk klasifikasi. Hasil dari peneltian tersebut menunjukkan bahwa algoritma *logistic regression* lebih unggul dibanding dengan algoritma SVM, dimana nilai akurasi yang diperoleh sebesar 76% untuk nilai *precision*, 71% untuk nilai *recall* dan 72% untuk nilai *F-Measure*. Sedangkan apabila menggunakan SVM saja hanya mendapatkan nilai *precision* sebesar 63%, nilai *recall* sebesar 60% dan nilai *F-measure* sebesar 62%. Sehingga dengan menerapkan metode WVSVM diharapkan dapat meningkatkan nilai akurasi pada klasifikasi halaman web, agar anak-anak dapat menelusuri internet dengan aman sesuai dengan usia mereka walaupun tanpa pengawasan orang tua.

2. Metode Penelitian

Tahapan yang dilakukan dalam penelitian ini dapat dilihat pada gambar



Gambar 2.1 Alur sistem klasifikasi halaman web anak

Adapun tahapan yang yang akan dilakukan dalam penelitian ini adalah sebagai berikut :

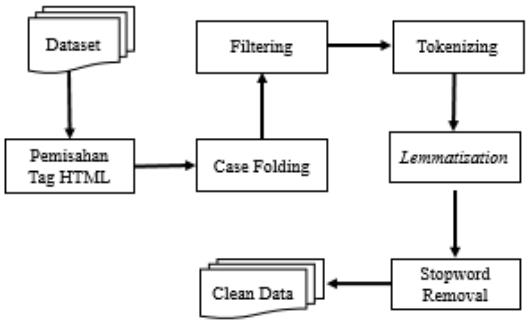
2.1. Crawling Halaman Web

Proses *crawling* pada halaman web http://dmoz-odp.org/ untuk mendapatkan menyalin semua informasi yang dibutuhkan sebagai data untuk proses klasifikasi halaman web. Penulis mengambil daftar halaman web yang ditujukan untuk anak dengan mengacu pada kategori *Kids & Teens Directory.* Data yang bersifat positif diambil dari halaman web yang bertanda '[*kids*]'. Sedangkan untuk data negatif yaitu halaman web yang ditujukan untuk umum penulis menggunakan kumpulan halaman web yang berada pada kategori *arts, business, computers, games, health, home, news, recreation, reference, regional, science, shopping, society, sport*. Jumlah data yang digunakan sebagai data set berjumlah adalah 3302 data positif dan 3516 data negative.

2.2. Tahap Preprocessing

Data yang digunakan dalam penelitian ini adalah kumpulan website yang telah di *crawling* dari website http://dmoz-odp.org.

Terdapat 5 tahap preporcessing[6]., yaitu:



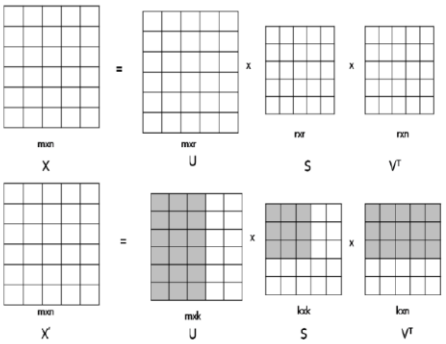
Gambar 2.2. Tahap preprocessing

1. ***Case Folding***
   Case Folding yaitu mengubah semua huruf dalam teks menjadi huruf kecil.
2. ***Filtering***
   Filtering yaitu menghapus semua karakter kecuali huruf a-z.
3. ***Tokenizing***
   Tokenizing adalah sebuah proses untuk memilah isi teks sehingga menjadi satuan kata-kata.
4. ***Lemmatization***
   Lemmatization merupakan suatu proses untuk mengubah kata ke bentuk dasarnya.
5. ***Stopword Removal***
   Stopword removal adalah tahap untuk menghilangkan kata yang tidak penting seperti: saya, adalah, yang, dan sebagainya.

2.3. Ekstraksi Fitur

Setelah data melewati tahap preprocessing, selanjutnya data akan melalui tahap ekstraksi fitur dengan menggunakan 2 algoritma, yaitu *Latent semantic analysis* dan *web page feature selection*, untuk klasifikasi halaman web menggunakan algoritma SVM.

1. ***Latent Semantic Analysis***
   Latent Semantic Analysis (LSA) merupakan sebuah metode statistik yang dapat digunakan untuk menentukan dan merepresentasikan kesamaan makna dari kata-kata dan teks dengan cara melakukan analisis terhadap teks dalam jumlah yang besar[7].
   Teknik SVD yang digunakan pada LSA adalah *reduced* SVD, yaitu akan dilakukan proses pengurangan dimensi pada matriks hasil dekomposisi SVD.



Gambar 2.3. SVD Matriks

Setiap matriks, misalnya matriks berukuran $t \times d$ yang direpresentasikan sebagai **X**, seperti matriks term x dokumen dapat didekomposisi ke dalam bentuk persamaan[8]:

$$X = U \times S \times V^T \quad (1)$$

Setelah diproses menggunakan SVD, persamaan (1) akan disedernakan menjadi:

$$X = U_k \times S_k \times V_k{}^T \quad (2)$$

Dimana U merupakan matriks dari vektor dokumen, V^T merupakan matriks dari vektor *term* dan S merupakan matriks diagonal yang berisikan nilai-nilai singular. Karena dalam penelitian ini ingin menemukan hubungan semantik yang sama antara dokumen yang berbeda, maka vektor matriks yang digunakan hanya vektor dokumen. Oleh karena itu, dalam penelitian ini akan menggunakan persamaan (3) untuk mendapatkan vektor semantik dari setiap dokumen

$$X = U_k \times S_k \quad (3)$$

2. *Web Page Feature Selection*

Fitur halaman web dapat digunakan untuk menilai kategori halaman web yang diberikan. Dalam penelitian ini, penulis mengekstraksi 4 fitur teks yang berbeda[4], yaitu:

a. Frekuensi kata kunci
   Untuk menghitung frekuensi kata kunci dapat menggunakan notasi berikut:

$$\sum_{k=1}^{n} \text{term}_k \quad (4)$$

b. Jumlah total kata yang ditampilkan dalam dokumen ‘x’

   Jumlah total kata : *W*

c. Rasio jumlah kata kunci terhadap total jumlah kata dalam dokumen.
   Untuk menghitung rasio jumlah kata kunci tersebut dapat menggunakan notasi berikut:

$$\text{Ratio} : \frac{\sum_{k}^{n} \text{term}_k}{W} \quad (5)$$

   Jika rasionya tinggi, maka konten halaman web mungkin termasuk kategori yang mengandung kata kunci tersebut.

d. Interval rata-rata antar istilah

$$\text{Average interval} : \frac{I_k}{\sum_{k}^{n} \text{term}_k} \quad (6)$$

Dimana $term_k$ adalah angka yang mewakili frekuensi $term_k$. L_k adalah interval antara istilah *term* ke-berapa.

Keyword yang digunakan pada ekstraksi fitur dengan menggunakan WPFS didapat dari mengekstrak seluruh teks yang berada dalam tag <meta> dengan atribut “class” bernilai “keyword” dari semua halaman web yang termasuk kedalam kategori ának. Kemudian keyword tersebut akan melalui *preprocessing* sehingga mendapatkan *term* yang bersifat unik. Sepuluh keyword dengan frekuensi paling banyak adalah kid (470), game (427), child (413), story (330), book (302), camp (235), school (235), dance (232), butterfly (183), dan science (157).

2.4. Klasifikasi SVM (*Support Vector Machine)*

Algoritma SVM pada dasarnya digunakan untuk proses klasifikasi antara dua kelas atau *binary classification*, seiring dengan perkembangannya SVM digunakan juga untuk klasifikasi *multi-class* yaitu dengan cara kombinasi antara beberapa *binary classifier*[9]. Fungsi keputusan pada SVM memanfaatkan sebuah fungsi kernel $K(x_i, x_d)$. Adapun beberapa fungsi kernel yang umum digunakan dapat dilihat pada persamaan (11), (12), (13), dan (14) berikut[8].

Linear :

$$K(x_i, x_j) = x_i^T x_j \quad (11)$$

Polinomial :

$$K(x_i, x_j) = (\gamma . x_i^T x_j + r)^d, \gamma > 0 \quad (12)$$

Radial basis function (RBF)

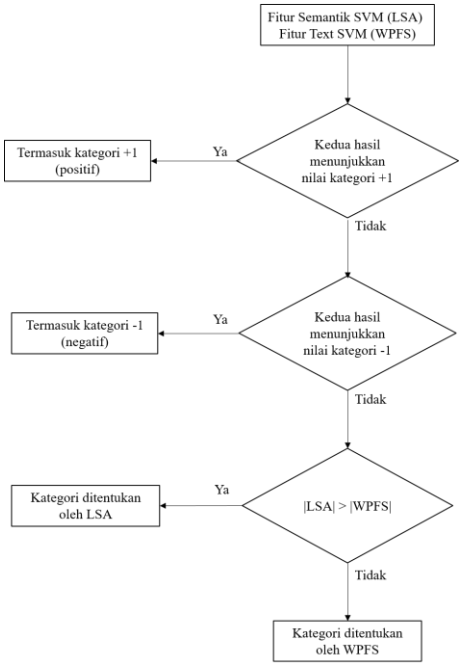
$$K(x_i, x_j) = \exp\left(-\gamma \|x_i - x_j\|^2\right), \gamma > 0 \quad (13)$$

Sigmoid

$$K(x_i, x_j) = \tanh \gamma\, x_i^T x_j + r \quad (14)$$

2.5. *Voting* Kategori

Pada penelitian ini akan menghasilkan dua buah model SVM yang berbeda, yaitu dengan seleksi fitur LSA dan WPFS. Apabila model klasifikasi yang dihasilkan dari LSA dan WPFS bernilai sama yaitu +1 maupun -1 maka data tersebut termasuk kedalam kategori +1 atau -1 (tanpa melalui proses voting). Sedangkan apabila nilai yang dihasilkan oleh model dengan LSA dan WPFS berbeda, maka diperlukan proses *voting* untuk menentukan kategori dengan mengacu pada nilai LSA atau WPFS yang paling besar. Untuk menentukan kategori dari data latih tersebut penulis menggunakan skema *voting* seperti gambar 2.2 berikut:



Gambar 2.4. Alur proses voting

2.6. Pengujian dan Analisa hasil program

Pada penelitian ini, penulis akan menguji performa SVM dengan menggunakan kernel linier, kernel polinomial, kernel RBF dan kernel dalam dua ratio dataset yang berbeda yaitu 80:20 dan 70:30 untuk menemukan kernel yang memberikan performa klasifikasi dengan algoritma WVSVM yang terbaik[10]. Hasil performa dengan kernel terbaik tersebut akan dibandingkan dengan performa dari hasil pengujian LSA-SVM dan WPFS-SVM menggunakan kernel yang sama untuk melihat seberapa baik metode Weighted Voting Support Vector Machine untuk klasifikasi halaman web anak. Pengujian dilakukan dengan menggunakan *confussion matrix* sebagai berikut:

Table 2.1. Confussion matrix

		Aktual	
	Class	Positive	Negative
Prediksi	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

Untuk menghitung akurasi, digunakan persamaan sebagai berikut:

$$Akurasi = \frac{N_{benar}}{N} \times 100\% \quad (15)$$

Untuk menghitung F-measure adalah sebagai berikut:

$$F\ measure = 2 \times \frac{presisi \times recall}{(presisi + recall)} \quad (16)$$

Sedangkan untuk menghitung nilai *presisi* dan *recall* dapat menggunakan rumus berikut:

$$Presisi = \frac{TP}{TP+FP} \quad (17)$$

$$Recall = \frac{TP}{TP+FN} \quad (18)$$

3. Hasil Penelitian dan Pembahasan

Proses pengujian ini dilakukan dalam 2 skenario pengujian. Skenario pengujian pertama akan membandingkan nilai performa dari dua ratio yang berbeda yakni 70:30 dan 80:20 untuk mengetahui ratio dan kernel terbaik yang memberikan performa terbaik pada klasifikasi dengan algoritma WVSVM, selanjutnya pada skenario pengujian kedua ratio dan kernel terbaik akan diterapkan pada LSA-SVM dan WPFS-SVM untuk membandingkan performa hasil pengujian dari tiga algoritma tersebut.

3.1. Skenario Pengujian I

A. Ratio data 70:30

Untuk ratio 70% data latih dan 30% data uji hasil pengujian dengan menggunakan confussion matrix dapat dilihat pada tabel dibawah ini:

Tabel 3.1. Hasil Pengujian WVSVM dengan dataset 70:30

Kernel	Akurasi	Presisi	Recall	F-measure
Linear	66%	64%	**65%**	65%
Polynomial	60%	**85%**	19%	31%
RBF	**71%**	72%	**65%**	**68%**
Sigmoid	53%	51%	50%	51%

Dari hasil pengujian pada tabel 4.3 dengan perbandingan data train dan data tes sebanyak 70:30, hasil akurasi tertinggi diperoleh dengan menggunakan kernel RBF dengan nilai 71%.

B. Ratio data 80:20

Pengujian kedua pada skenario pertama membagi data menjadi 80% data train dan 20% data test. Hasil pengujian yang didapatkan adalah sebagai berikut:

Tabel 3.2. Hasil Pengujian WVSVM dengan dataset 80:20

Kernel	Akurasi	Presisi	Recall	F-measure
Linear	72%	72%	**68%**	70%
Polynomial	59%	**85%**	19%	31%
RBF	**74%**	76%	**68%**	**71%**
Sigmoid	54%	52%	53%	52%

Dari hasil pengujian pada tabel 3.2 dengan perbandingan data train dan data tes sebanyak 80:20, hasil akurasi tertinggi diperoleh dengan menggunakan kernel RBF dengan nilai 74%. Dari pengujian dengan kedua ratio tersebut dapat diketahui bahwa WVSVM memiliki performa terbaik dengan ratio data 80:20 dan kernel RBF

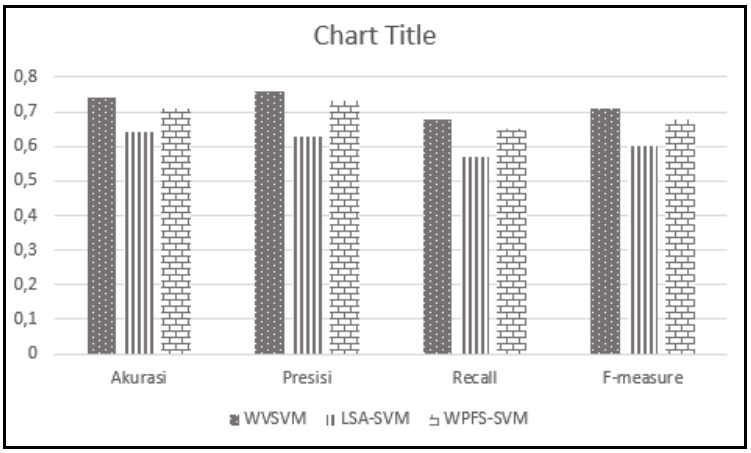
3.2. Skenario Pengujian II

Dari dua ratio data set pada pengujian pertama, didapatkan ratio terbaik untuk sistem klasifikasi dengan menggunakan WVSVM adalah sebesar 80:20 dan menggunakan kernel RBF, dengan menggunakan ratio dan kernel yang sama penulis akan membandingkan performa WVSVM dengan LSA-SVM dan WPFS-SVM. Hasil pengujian dari ketiga algoritma tersebut dapat dilihat pada tabel 3.3. berikut:

Tabel 3.3.. Hasil pengujian klasifikasi halaman web anak

Kernel	Akurasi	Presisi	Recall	F-measure
WVSVM	**74%**	**76%**	**68%**	**71%**
LSA-SVM	64%	63%	57%	60%
WPFS-SVM	71%	73%	65%	68%

Hasil pengujian pada tabel 3.1 direpresentasikan pada grafik dibawah ini:



Gambar 3.1. Hasil pengujian klasifikasi halaman web anak

Dari dua skenario pengujian diatas dapat diketahui bahwa WVSVM memiliki performa pengujian lebih baik dibandingkan dengan algoritma LSA-SVM dan WPFS-SVM dengan nilai akurasi sebesar 74%, nilai presisi sebesar 76%, nilai *recall* sebesar 68%, dan nilai 71% untuk *F-measure* dengan ratio 80% data latih dan 20% data uji menggunakan kernel RBF. Sedangkan algoritma yang memiliki performa pengujian paling rendah adalah LSA-SVM yaitu 64% untuk akurasi, 63% untuk nilai presisi, 57% untuk nilai *recall* dan 60% untuk nilai *F-measure.*

4. Kesimpulan

4.1. Kesimpulan

Berdasarkan hasil penelitian klasifikasi halaman web anak menggunakan algoritma *Weighted Voting Support Vector Machine* dapat disimpulkan bahwa hasil pengujian terbaik didapatkan dengan menerapkan ratio data latih dan data uji sebesar 80:20 serta menggunakan kernel RBF dengan nilai akurasi sebesar 74%, dengan nilai presisi sebesar 76%, nilai recall sebesar 68% dan nilai f-measure sebesar 71%. Performa pengujian WVSVM lebih unggul dibandingkan dengan WPFS-SVM yang mendapatkan akurasi sebesar 71%, presisi sebesar 73%, *recall* sebesar 65% dan *F-measure* sebesar 68% serta LSA-SVM yang mendapatkan nilai 64% untuk akurasi, 63% untuk presisi, 57% untuk *recall* dan 60% untuk *F-measure.*

4.2. Saran

Saran yang dapat diberikan terhadap pengembangan sistem klasifikasi halaman web untuk anak dengan menggunakan metode WVSVM adalah sebagai berikut:

1. Pada tahap preprocessing dilakukan n-gram karena ada teks hasil crawling tidak dipisahkan dengan spasi sehingga menjadi satu kata yang sangat panjang.
2. Menggunakan algoritma ekstraksi fitur yang lain agar hasil pengujiannya dapat lebih baik.

Refrensi

[1] A. Sun, E.-P. Lim, And W.-K. Ng, “Web Classification Using Support Vector Machine,” *Proc. Fourth Int. Work. Web Inf. Data Manag. Widm 02*, 2002.

[2] S. Dixit And R. K. Gupta, “Layered Approach To Classify Web Pages Using Firefly Feature Selection By Support Vector Machine (Svm),” *Int. J. U- E-Service, Sci. Technol.*, 2015.

[3] K. He And C. Li, “Structure-Based Classification Of Web Documents Using Support Vector Machine,” *Proc. 2016 4th Ieee Int. Conf. Cloud Comput. Intell. Syst. Ccis 2016*, Pp. 215–219, 2016.

[4] R. C. Chen And C. H. Hsieh, “Web Page Classification Based On A Support Vector Machine Using A Weighted Vote Schema,” *Expert Syst. Appl.*, Vol. 31, No. 2, Pp. 427–435, 2006.

[5] P. Serdyukov, “Web Page Classification On Child Suitability,” No. January, 2010.

[6] A. Setiawan, I. F. Astuti, And A. H. Kridalaksana, “Klasifikasi Dan Pencarian Buku Referensi Akademik Menggunakan Metode Naïve Bayes Classifier (Nbc) (Studi Kasus : Perpustakaan Daerah Provinsi Kalimantan Timur),” Vol. 10, No. 1, 2015.

[7] T. K. Landauer, P. W. Foltz, And D. Laham, “An Introduction To Latent Semantic Analysis,” *Discourse Process.*, 1998.

[8] R. Adhitia And A. Purwarianti, “Penilaian Esai Jawaban Bahasa Indonesia Menggunakan Metode Svm - Lsa Dengan Fitur Generik,” *J. Sist. Inf.*, Vol. 5, No. 1, P. 33, 2016.

[9] O. Somantri, D. Apriliani, J. T. Informatika, P. Harapan, And B. Tegal, “Support Vector Machine Berbasis Feature Selection Untuk Sentiment Analysis Kepuasan Pelanggan Terhadap Pelayanan Support Vector Machine Based On Feature Selection For Sentiment Analysis Customer Satisfaction On Culinary,” Vol. 5, No. 5, Pp. 537–548, 2018.

[10] S. Mayor And B. Pant, “Document Classification Using Support Vector Machine,” *2017 Int. Conf. Curr. Trends Comput. Electr. Electron. Commun.*, Vol. 4, No. 04, Pp. 1741–1745, 2012.