Jurnal TRANSFORMATIKA,
Vol.19, No.2, Januari 2022, pp. 144-150
P-ISSN: 1693-3656, E-ISSN: 2460-6731
https://journals.usm.ac.id/index.php/transformatika/

Content Classification based-on Latent Semantic Analysis and Support Vector Machine (LSA-SVM)

Gita Indah Marthasari[1*], Nur Hayatin[2], Maulidya Yuniarti[3]
[1,2,3]Universitas Muhammadiyah Malang
Jalan Tlogomas 246, Malang, Jawa Timur Indonesia,
e-mail: gita@umm.ac.id [1], noorhayatin@umm.ac.id [2], maulidyayuniarti@webmail.umm.ac.id [3]

ARTICLE INFO

History of the article :
Received 22 Oktober 2020
Received in revised form 23 Maret 2021
Accepted 27 Januari 2022
Available online 31 Januari 2022

Keywords:

web page classification, support vector machine, latent semantic analysis

*** Correspondece:**

Telepon:
+6281333863200

E-mail:
gita@umm.ac.id

ABSTRACT

The diversity of the content of a web page can have a negative impact if used by the wrong user. Almost a half of internet users are children. Therefore, it is important to classify web pages to find out which pages are worthy of being seen by children and that are not feasible. One method that can be used is the Support Vector Machine (SVM) algorithm. SVM is a binary classification whose working principle is to find the best hyperplane to separate the two classes. To obtain better classification accuracy, the SVM is combined with the Latent Semantic Analysis (LSA) algorithm. The data used in this study were taken from the DMOZ web data which has been classified into two categories. The data is then entered into the pre-processing stage for further feature extraction using LSA. The LSA algorithm is used to find out the semantic similarities of words and text contained in web pages. The results of feature extraction are then classified using SVM with RBF kernel. Based on the testing result, we obtain a classification accuracy of 64%.

INTRODUCTION

Menurut *InternetWorldStats* pada kuartal pertama tahun 2020 terdapat lebih dari 4 miliar manusia di seluruh dunia yang mengakses internet setidaknya sekali dalam satu bulan [1]. Pesatnya pertumbuhan pengguna internet tersebut turut andil dalam memicu pertumbuhan halaman web. Peningkatan jumlah halaman web ini terjadi baik dalam topik maupun informasi yang terdapat didalamnya. Namun, informasi tersebut dapat memberi dampak negatif pada beberapa proses antara lain pengembalian hasil pencarian yang tidak relevan dengan kata kunci yang telah diinputkan oleh pengguna pada mesin pencari.

Saat ini bukan hanya orang dewasa yang menjadi pengguna internet, tetapi juga anak usia sekolah. Penelitian UK menemukan lebih dari 40% anak usia 5-15 tahun menelusuri internet

tanpa pengawasan dari orang tua [2]. Padahal dengan pengembalian halaman web yang tidak relevan dapat menyebabkan anak-anak tersebut mengakses halaman web yang tidak sesuai dengan usia mereka. Hal ini dapat berpotensi membahayakan anak-anak mengingat banyaknya dampak negatif apabila internet tidak dipergunakan dengan hati-hati antara lain kejahatan yang bermula dari internet, pornografi, dan *cyber bullying*.

Mengacu pada latar belakang tersebut maka dibutuhkan sebuah sistem yang dapat melakukan klasifikasi halaman web secara otomatis untuk membedakan web yang layak diakses oleh anak dan tidak. Dengan adanya sistem tersebut tentunya juga dapat membuat halaman web yang diakses oleh anak-anak lebih dapat terkontrol sesuai dengan usia mereka. Klasifikasi halaman web dapat dilakukan dengan mengekstrak terms dan *tag* HTML yang terdapat pada halaman web.

Terdapat berbagai macam algoritma klasifikasi yang dapat diterapkan pada klasifikasi halaman web, yaitu *Decision Tree, NBC (Naive Bayes Classifier), Logistic Regression,* KNN (K-*Nearest Neighbor*) dan SVM (*Support Vector Machine*). Jika mengacu dari popularitas, maka algoritma SVM termasuk sebagai algoritma yang populer digunakan dalam proses klasifikasi [3]. Algoritma SVM termasuk dalam kategori algoritma *supervised machine learning* yang dapat digunakan untuk klasifikasi halaman web dengan melatih sistem dengan pola yang diketahui menggunakan tingkat pembelajaran yang telah ditentukan. Adanya variasi dalam pembelajaran memungkinkan tingkat akurasi yang diinginkan dapat dicapai menggunakan SVM. SVM merupakan metode yang efektif untuk mengklasifikasikan data berdimensi tinggi [4]. Klasifikasi dokumen menggunakan SVM telah dilakukan oleh [5] menggunakan data medis yang bertujuan mengklasifikasikan jenis kanker berdasarkan kode genetik.

Penelitian [6] juga memilih metode SVM untuk klasifikasi dokumen dengan jumlah dataset besar. Data yang digunakan dalam klasifikasi adalah data berita dari web BBC yang telah dikategorikan ke dalam 5 (lima) kelas yaitu bisnis, entertainment, teknologi, olahraga, dan politik sebanyak 2225 dokumen. Selain itu, dataset juga berasal dari 20 (dua puluh) kelompok berita dengan total dokumen sebanyak 19.999 data yang terbagi ke dalam 20 (dua puluh) kelas yang berbeda. Akurasi klasifikasi menggunakan SVM terhadap data uji mencapai 88% untuk dataset dari web BBC dan 94.5% untuk dataset dari kelompok berita.

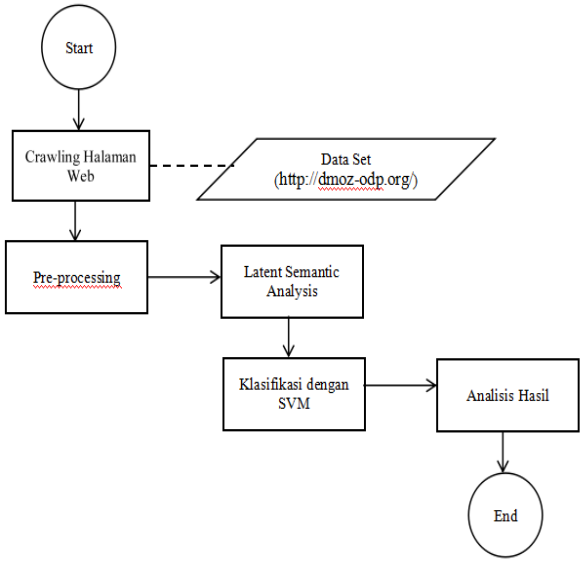
LSA (*Latent Semantic Analysis*) menggunakan metode aljabar linear dengan menerapkan *singular value decomposition* (SVD). LSA akan membentuk matriks yang merepresentasikan hubungan antara term-dokumen yang merupakan *semantic space* yakni kata-kata dan dokumen-dokumen yang berhubungan dekat akan dihubungkan satu sama lain. LSA digunakan untuk mengekstrak hubungan semantik umum antara kata kunci dengan dokumen. Pada [6] juga dilakukan percobaan mengkombinasikan SVM dengan LSA menggunakan 2 (dua) jenis dataset. Dari kedua dataset yang diuji, kombinasi SVM dan LSA mampu melakukan klasifikasi dengan efisiensi waktu yang lebih baik dibandingkan jika hanya menggunakan SVM.

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]. Penelitian lain yang mengkombingasikan SVM dan LSA adalah klasifikasi sentimen oleh [8]. Data yang digunakan adalah review online yang telah diberi label berdasarkan empat jenis emosi yaitu *happiness, hope, disgust,* dan *anxiety*. Penelitian ini bertujuan melakukan klasifikasi halaman web anak dengan menerapkan algoritma *Support Vector Machine* yang dikombinasikan dengan *Latent Semantic Analysis.* Penggabung SVM dan LSA meningkatkan efektivitas klasifikasi terutama ketika data yang digunakan berjumlah besar.

RESEARCH METHODS

Untuk mengklasifikasi halaman web anak menggunakan metode LSA-SVM, ada 4 tahapan proses yang dilakukan (Gambar 1), yaitu : *crawling, preprocessing*, ekstraksi fitur dengan algoritma LSA, dan klasifikasi dengan algoritma SVM. Adapun dataset yang digunakan dalam

penelitian ini diambil dari direktori web DMOZ (*http://dmoz-odp.org/*). Halaman web yang diambil mengacu pada kategori *Kids & Teens Directory* (Direktori anak dan remaja awal, dalam Bahasa Indonesia). Data yang bersifat positif diambil dari halaman web yang bertanda '[*kids*]'. Sedangkan untuk data negatif yaitu halaman web yang ditujukan untuk umum (selain anak) menggunakan kumpulan halaman web yang berada pada kategori *Home*. Jumlah data yang digunakan dalam penelitian adalah sebanyak 5315 data yang terbagi menjadi 2562 data positif dan 2753 data negatif.

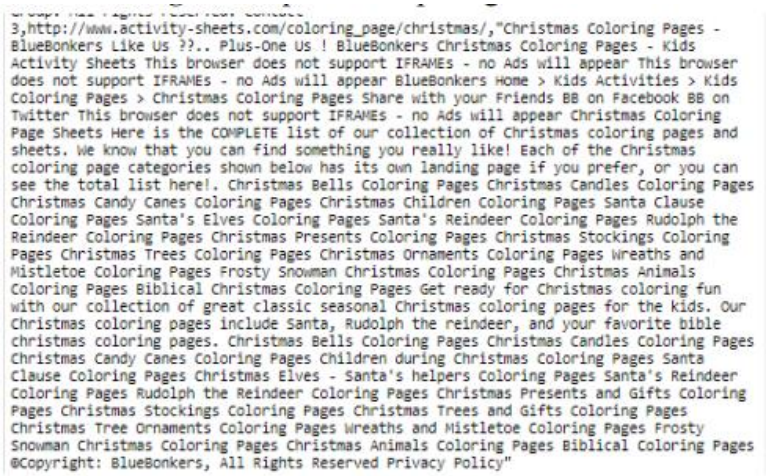


Gambar 1. *Flowchart* Klasifikasi Halaman Web Anak Dengan Metode LSA-SVM

a. Crawling Halaman Web

Pada tahap ini penulis melakukan proses *crawling* pada halaman web http://dmoz-odp.org/ untuk menyalin semua informasi yang dibutuhkan sebagai data untuk proses klasifikasi halaman web. *Crawling* adalah proses otomasi dalam navigasi dan mendapatkan halaman web menggunakan software atau alat tertentu [9]. Proses crawling ini akan menyalin semua informasi seperti HTML yang ada didalam halaman web tersebut sehingga dapat digunakan sebagai data acuan pada proses klasifikasi. Dokumen HTML lebih dari sekedar file teks sederhana, HTML bersifat terstruktur dan terhubung dengan dokumen HTML yang lainnya. Proses crawling menggunakan WebHarvy, adapun contoh hasil crawling dapat dilihat pada gambar 2.

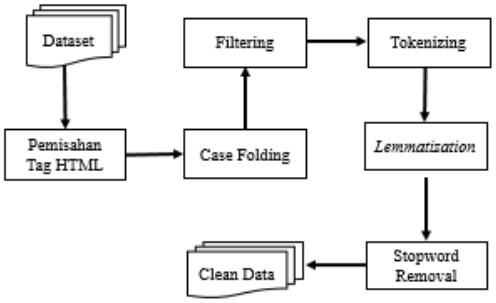
Gambar 2. Hasil crawling halaman web



b. Preprocessing

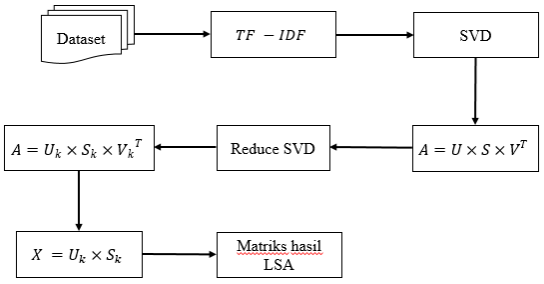
Preprocessing yang dilakukan dalam penelitian ini terdiri dari 5 tahap, meliputi : pemisahan HTML Tag, *Case Folding* (mengubah semua huruf dalam teks menjadi huruf kecil), *filtering* (mengubah semua huruf dalam teks menjadi huruf kecil), *Tokenizing* (sebuah proses untuk memilah isi teks sehingga menjadi satuan kata-kata), *Lemmatization* (suatu proses untuk mengubah kata ke bentuk dasarnya), dan *Stopword Removal* (tahap untuk menghilangkan kata yang tidak penting seperti: saya, adalah, yang, dan sebagainya), seperti yang ditunjukkan pada Gambar 3.

Gambar 3. Tahapan Preprocessing Pada Proses Klasifikasi Halaman Web



c. Ekstraksi Fitur dengan LSA

Setelah data melewati tahap preprocessing, selanjutnya data akan melalui tahap ekstraksi fitur dengan menggunakan algoritma LSA sebelum dilakukan proses klasifikasi. Algoritma LSA digunakan untuk mengetahui kesamaan semantik dari kata-kata dan teks yang ada didalam halaman web. Secara garis besar alur algoritma LSA ditunjukkan pada Gambar 4.



Gambar 4. Tahapan Ekstraksi Fitur menggunakan LSA

LSA menggunakan metode aljabar linear *Singular Value Decomposition* (SVD) yang akan mendekomposisikan matriks *A* (*term-documen*t) menjadi matriks *U, S,* dan *VT*. Tahap pertama dalam LSA adalah mengubah data kedalam bentuk term-document matriks, dimana pada penelitian ini akan menggunakan TF-IDF. Selanjutnya Matriks TF-IDF akan membentuk sebuah matriks dengan term sebagai kolom, dan document sebagai baris. Tahap selanjutnya adalah melakukan dekomposisi matriks A dengan menggunakan SVD. Teknik SVD yang digunakan dalam LSA adalah reduced SVD, dimana akan dilakukan proses pengurangan dimensi dengan cara memotong matriks tersebut dengan menentukan nilai *k* pada matriks diagonal yang berisi nilai-nilai singular. Setiap matriks, misalnya matriks berukuran $t \times d$ yang direpresentasikan sebagai X, seperti matriks term *x* dokumen dapat didekomposisi ke dalam bentuk persamaan [10].

LSA atau penggurangan dimensi menggunakan pemotongan SVD pada penelitian ini

DOI : 10.26623/transformatika.v19i2.2745

menggunakan modul '*truncatedSVD'* yang disediakan oleh library *sklearn*. Data yang digunakan pada ekstraksi fitur LSA adalah data teks dari kolom "Data" SVD atau di representasikan dalam bentuk matriks *X=U_k×S_k×(V_k)^T.* Matriks *U* merupakan matriks dokumen – term yang menggambarkan keterkaitan semantic antar dokumen berbentuk *m x r, S* merupakan matriks diagonal berisi nilai scalar berbentuk *r x r*, dan *VT* merupakan matriks term-topik yang berbentuk r x n. Dalam penelitian ini matriks yang digunakan sebagai data latih yang akan diterapkan kedalam model adalah matriks *U_k×S_k.* LSA diterapkan pada data latih dan data uji. Pada penelitian ini, matriks yang digunakan sebagai data pada penelitian ini merupakan hasil perkalian dari matriks *U* dan *S* yang akan menghasilkan nilai semantik antar dokumen.

d. Klasifikasi menggunakan Metode Support Vector Machine

Setelah melalui tahap ekstraksi fitur, selanjutnya halaman web akan melalui tahap klasifikasi menggunakan algoritma SVM. Algoritma SVM adalah suatu teknik untuk melakukan prediksi, baik dalam kasus klasifikasi maupun regresi [11]. Teknik ini termasuk dalam metode klasifikasi supervised machine learning karena memiliki target pembelajaran tertentu.

Algoritma SVM pada dasarnya digunakan untuk proses klasifikasi antara dua kelas atau *binary classification.* Seiring perkembangannya, SVM juga digunakan untuk klasifikasi multi-class yaitu dengan cara kombinasi antara beberapa *binary classifier*. Fungsi keputusan pada SVM memanfaatkan sebuah fungsi kernel. Adapun beberapa fungsi kernel yang umum digunakan adalah linear, polynomial, radial basis function, dan sigmoid [12]. Persamaan keempat kernel tersebut dapat dilihat pada (1), (2), (3), dan (4) dimana *x* merupakan vector input, while *v* is eigenvectors.

Linear :

$$K(x_i, x_j) = (x_i \cdot x_j); \quad (1)$$

Polinomial :

$$K(x_i, x_j) = (x_i \cdot x_j + 1)^p; \quad (2)$$

Radial basis function (RBF) :

$$K(x_i, x_j) = e^{-\gamma(x_i - x_j)^2}; \quad (3)$$

Sigmoid :

$$K(x_i, x_j) = tanh(\eta x_i \cdot x_j + v); \quad (4)$$

RESULTS AND DISCUSSION

Proses build model dengan menggunakan LSA akan dilakukan dengan menggunakan kernel RBF. Pada penelitian ini, digunakan 2 (dua) skenario pengujian dengan komposisi data latih dan data uji yang berbeda yaitu 70:30 dan 80:20. Selanjutnya, akan diuji performa dari LSA-SVM menggunakan *confussion matrix* untuk mengukur seberapa baik metode yang diusulkan dalam melakukan klasifikasi halaman web anak. Dari hasil pengujian nanti akan didapatkan nilai precision, recall, dan f-measure dari implementasi LSM-SVM. Tabel 1 berikut adalah hasil pengujian yang telah dilakukan pada penelitian ini.

Tabel 1 Hasil Pengujian LSM-VSM

Precision	Recall	F-measure
64%	63%	57%

Selain itu juga dilakukan perbandingan metode LSM-VSM dengan baseline SVM [6] dengan menghitung nilai akurai menggunakan persamaan (5) dimana N_{benar} merupakan jumlah prediksi benar dan N adalah jumlah data.

$$Akurasi = \frac{N_{benar}}{N} \times 100\% \quad (5)$$

Berdasarkan pengujian, komposisi data latih dan data uji yang terbaik adalah 80:20 dengan akurasi sebesar 64%. Namun, dibandingkan dengan [6], akurasi klasifikasi menggunakan kombinasi SVM-LSA memiliki nilai akurasi lebih rendah (lihat Tabel 2). Hal ini karena LSA melakukan ekstraksi fitur sehingga dimungkinkan menghapus kata kunci penting yang mengindikasikan konten web anak. Oleh karena itu, perlu dilakukan eksperimen pada perbaikan metode LSA atau dapat menggunakan metode ekstraksi fitur yang lain. Namun, LSA memungkinkan berkurangnya dimensi dataset sehingga meningkatkan efisiensi proses klasifikasi [6].

Tabel 2 Perbandingan Akurasi Klasifikasi menggunakan SVM dan LSA-SVM

	SVM	LSA-SVM
Akurasi	88%	64%

CONCLUSIONS AND RECOMMENDATIONS

Penelitian ini mengusulkan metode LSA-SVM untuk mengklasifikasi halaman web anak menggunakan konten web berdasarkan tag HTML. Berdasarkan hasil pengujian, klasifikasi halaman web anak menggunakan metode LSA-SVM didapatkan nilai akurasi sebesar 64% dimana nilai ini lebih rendah dari klasifikasi menggunakan SVM. Hal ini dikarenakan LSA mereduksi dimensi data, sehingga dimungkinkan banyak data representative yang akan hilang yang dapat mempengaruhi performansi hasil klasifikasi, terutama jika data yang digunakan kecil. Meski demikian, kombinasi LSA-SVM terbukti meningkatkan efisiensi waktu klasifikasi karena berkurangnya dimensi data set. Oleh karena itu, jika klasifikasi menitik-beratkan pada akurasi, maka SVM sebaiknya tidak dikombinasikan dengan LSA.

Ucapan Terima Kasih

Terima kasih disampaikan kepada Direktorat Penelitian dan Pengabdian Masyarakat Universitas Muhammadiyah Malang sebagai pihak yang memberikan dukungan baik dalam hal akademik maupun administratif.

REFERENCES

[1] InternetWorldStats, “Internet World Stats : Usage and Population Statistics,” *Miniwatts Marketing Group*, 2020. .

[2] S. Vashinda and J. Pilgrim, “Are we preparing students for the web in the wild? An analysis of features of websites for children,” *J. Lit. Technol.*, vol. 20, no. 2, pp. 97–124, 2019, [Online]. Available: http://proxy.libraries.smu.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=eue&AN=137069217&site=ehost-live&scope=site.

[3] W. Huang and H. You, “Web Page Classification Algorithm Based on Semi-Supervised Support Vector Machine,” in *Proceedings of 2018 2nd IEEE Advanced Information*

Management, Communicates, Electronic and Automation Control Conference, IMCEC 2018, 2018, no. Imcec, pp. 2144–2148, doi: 10.1109/IMCEC.2018.8469700.

[4] S. Shinde, P. Joeg, and S. Vanjale, "Web Document Classification using Support Vector Machine," in *International Conference on Current Trends in Computer, Electrical, Electronics and Communication, CTCEEC 2017*, 2018, pp. 688–691, doi: 10.1109/CTCEEC.2017.8455102.

[5] B. Ghaddar and J. Naoum-Sawaya, "High dimensional data classification and feature selection using support vector machines," *Eur. J. Oper. Res.*, vol. 265, no. 3, pp. 993–1004, 2018, doi: 10.1016/j.ejor.2017.08.040.

[6] V. Khatavkar and P. Kulkarni, *Comparison of support vector machines with and without latent semantic analysis for document classification*, vol. 808. Springer Singapore, 2019.

[7] L. Cagliero, P. Garza, and E. Baralis, "ELSA: A multilingual document summarization algorithm based on frequent itemsets and latent semantic analysis," *ACM Trans. Inf. Syst.*, vol. 37, no. 2, pp. 1–33, 2019, doi: 10.1145/3298987.

[8] W. Zhang, S. xi Kong, Y. chun Zhu, and X. le Wang, "Sentiment classification and computing for online reviews by a hybrid SVM and LSA based approach," *Cluster Comput.*, vol. 22, pp. 12619–12632, 2019, doi: 10.1007/s10586-017-1693-7.

[9] I. Hernández, C. R. Rivero, and D. Ruiz, *Deep Web crawling: a survey*, vol. 22, no. 4. 2019.

[10] M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," vol. 349, no. 6245, 2015, doi: 10.1126/science.aaa8415.

[11] 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)," *Inform. Mulawarman J. Ilm. Ilmu Komput.*, vol. 10, no. 1, p. 1, 2016, doi: 10.30872/jim.v10i1.17.

[12] J. Cervantes, F. Garcia-Lamont, L. Rodríguez-Mazahua, and A. Lopez, "A comprehensive survey on support vector machine classification: Applications, challenges and trends," *Neurocomputing*, no. xxxx, 2020, doi: 10.1016/j.neucom.2019.10.118.