OPEN ACCESS
J. OF INISTA, VOL. 2, NO. 1, PP.056-064, NOV 2019

ISSN : 2622-8106 (ONLINE)
HTTPS://DOI.ORG/10.20895/INISTA.V2I1

Journal of Informatics, Information System, SoftwareEngineeringandApplications

KlasifikasiAnalisis Sentimen Pada Gambar Meme Politik Dengan Library Tesseract Dan Algoritme Support Vector Machine

Eko Sanjaya[1)], Agi Prasetiadi[,2)] Wahyu Andi Saputra[,3)]

[1,2,3)]*Program Studi Teknik Informatika, Fakultas Teknik Industri dan Informatika,*
Institut Teknologi Telkom Purwokerto
Jl. D.IPanjaitan, No 128, Karangreja, Purwokerto Kidul, Kabupaten Banyumas, Jawa Tengah

Accepted on 12-08-2019

Abstract

Meme merupakan penyebaran informasi dalam bentuk gambar. Berdasarkan data yang diperoleh, pengembangan meme mulai meningkat menjelang pemilu 2019. Informasi yang diperoleh dari meme politik beragam. Salah satunya memberikan dukungan untuk suatu partai atau tokoh politik atau digunakan untuk mengkritik / mencaci-maki partai politik atau tokoh. Sehingga diperlukan suatu sistem yang dapat mengklasifikasikan meme berdasarkan kelas Penelitian ini bertujuan untuk menciptakan sistem yang dapat mengklasifikasikan meme politik berdasarkan kelas. Algoritma yang akan digunakan dalam mengklasifikasikan adalah Support vector macine (SVM) dengan ekstraksi fitur TF-IDF. Library yang akan digunakan dalam optical character recognition (OCR) adalah Tesseract. Berdasarkan hasil pengujian diketahui bahwa akurasi yang dihasilkan oleh SVM linier lebih baik daripada SVM non-linear. Akurasi terbaik dalam SVM linear dengan kombinasi TF-IDF adalah 75.71%.

Kata Kunci: Analisis sentimen, Meme politik, Ocrtesseract, Support Vector Machine

I. INTRODUCTION

Perkembangan teknologi dan informasi yang sangat pesat, memudahkan masyarakat berkomunikasi dengan yang lainnya. Perkembangan teknologi informasi ini, tidak luput dari perkembangan internet yang terus meningkat. Adanya internet, setiap manusia memiliki akses untuk dapat mencari dan memberikan informasi yang dibutuhkan. Dalam penyebaran informasi di social media mulai beragam, salah satunya menyebarkan informasi dalam bentuk teks gambar yang juga disebut *meme*. Istilah *meme* pertama kali dipopulerkan oleh Richard Dawkins dalam bukunya The Selfish Gene(1976). Menurut pemahaman Dawkins, *meme* adalah bentuk dari gen kebudayaan (ide, gagasan, pola perilaku, dan sebagainya) yang menyebar melalui proses imitasi, seperti halnya lagu, jargon, mode pakaian, hingga cara membangun gedung[1].

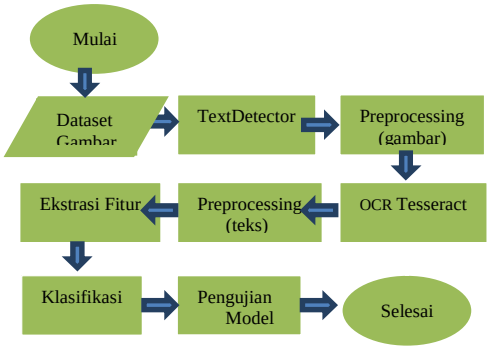
Penyampain informasi menggunakan *meme* mulai popular dikalangan masyarakat dalam belakangan ini. Kepopuleran tersebut ditunjukan dengan adanya website-website *meme* seperti 1cak.com dan *meme*comic.id. Dengan adanya website tersebut masyarat sering menggunakan *meme* di akun sosial media. *Meme* dapat dikategorikan menjadi dua[2], pertama *meme* ringan seperti *meme* candaan, motivasi, dan olahraga. Sedangkan *meme* kedua adalah *meme* yang berat seperti membahas pemerintahan dan politik. Kelebihan *meme* dalam penyampain informasi adalah menambahkan daya tarik masyarakat dalam membaca dan dapat mempermudah dalam menangkap informasi. Sekarang ini *meme* berat mulai sering digunakan oleh pendukung suatu partai politik untuk mendukung suatu partai atau tokoh politik yang didukung ataupun digunakan untuk mengkritik/ mencaci-makisuatu partai politik lain atau tokoh lain karena rasa ketidaksukaannya.

Dalam penggalian data dapat menggunakan *Text Mining*. Dengan menggunakan *Text Mining*, peneliti mendapatkan informasi yang berguna dan dapat meningkatkan pengetahuan bagi peneliti. *Text mining* dapat diterapkan dalam beberapa bidang, salah satunya adalah Analisis Sentimen. Analisis sentimen adalah bidang studi yang menganalisis pendapat seseorang, sentimen seseorang, evaluasi seseorang, sikap seseorang dan emosi seseorang ke dalam bahasa tertulis[3].

Text Mining juga dapat berfungsi sebagai klasifikasi, yang prosesnya untuk menemukan model untuk menggambarkan class atau konsep dari suatu data, dan juga digunakan untuk mendiskripsikan data penting serta dapat menprediksi data pada masa depan. Pada penelitian ini, data yang akan dianalisis adalah data berupa text yang diambil dari gambar *meme* yang didapatkan dari media sosial, khususnya *meme* mengenai politik.

Proses klasifikasi *meme* dapat dilakukan dengan menggunakan *Library* Tesseract untuk mengubah teks gambar/*meme* menjadi teks digital dan algoritma *Support vector machine* (SVM) digunakan untuk mengklasifikasi teks menjadi tigas kelas yaitu, kelas *meme* positif, kelas *meme* netral dan kelas positif. Tujuan dari penelitian ini adalah mendapatkan nilai akurasi menggunakan *Library*Tesseract danAlgoritme*Support Vector Machine* (SVM).

II.RESEARCH METHOD



Gambar 1. Tahapan Penelitian

A. Pengumpulan Dataset

Pada tahap ini pengumpulan dataset didapatkan dari postingan atau komentar di social media (Facebook, instagram dan twitter). Gambar yang diambil memiliki ketentuan seperti gambar minimal berkurang 320x320 pixel, teks mudah dibaca oleh system, background dari karakter harus kontras atau tidak memiliki background. Dataset yang digunakan dalam penelitian ini adalah 444 gambar *meme* dan diambil 31 sampel gambar *meme* secara random untuk dilakukan pengujian data baru.

B. TextDetector

Textdetector adalah menentukan suatu keberadaan objek (teks) dan ruang lingkupnya serta lokasi didalam sebuah gambar[4]. Tahapan *text detector* dalam penelitian ini menggunakan penelitian yang telah dilakukan oleh Zhou Xinyu dkk yaitu system *EAST: An Enfficient and Accurate Text Detector*[5]. *Text detector EAST* memiliki kelemahan tidak dapat memotong teks secara sempurna, Sehingga penulis memodifikasi system *EAST* agar dapat memotong teks dengan sempurna.

Gambar 2. Text Detector

Gambar 3. Hasil Memotong Text Detector

C. *Preprocessing (Gambar)*

Preprocessing pada gambar menjadi salah satu tahapan terpenting yang bertujuan untuk menghilangkan noise dan menyederhanakan gambar agar mudah dikelola saat tahap OCR tesseract. *Preprocessing* pada gambar terdiri lima tahap, yaitu:

1. Grayscale: menyederhanakan gambar.
2. Blur: mengurangi noise dan mengurangi detail.
3. Threshold: mengatur jumlah derajat keabuan pada citra.
4. Erosion: mengikis batas-batas tepi objek dan menghapus semua pixel yang dekat dengan batas. Sehingga erosion dapat menghilang/menghapus noise kecil.
5. Dilation: mempertebal font yang telah dikikis oleh tahap erosion dan dapat meningkat ukuran teks sehingga dapat mempermudah dalam pengenalan karakter.

Gambar 4. HasilPreprocessing Pada Gambar

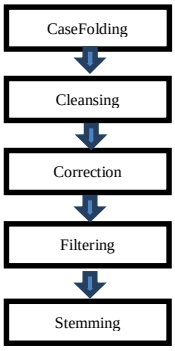
D. *OCR Tesseract*

OCR tesseract adalah tahapan mengekstrak teks didalam gambar menjadi teks digital dengan menggunakan library tesseract dengan akurasi tinggi. Tesseract dikembangkan sebagai perangkat lunak berpemilik oleh Hewlett Packard Labs. Pada tahun 2005, itu terbuka bersumber dari HP bekerja sama dengan University of Nevada, Las Vegas. Sejak 2006 telah dikembangkan secara aktif oleh Google dan banyak kontributor open source[6].

Gambar 5. Hasil OCR Tesseract

E. *Preprocessing(Teks)*

Preprocessing pada teks adalah tahapan memperbaiki teks hasil OCR tesseract yang belum sempurna dan juga berguna untuk mempersiapkan data teks menjadi lebih struktur. Tahapan *preprocessing* pada teks terdiri lima, yaitu:



Gambar 6. Tahapan Preprocessing

1. *CaseFolding:*merubahkarater menjadi bentuk standard(*uppercase* atau *lowercase*).
2. *Cleansing:* menghapus karakter seperti symbol, tanda Tanya dan angka.
3. *Correction:* memperbaiki ejaan yang tidak sesuai dengan KBBI tapi dapat diperbaiki dan menghapus ejaan yang tidak sesuai dengan KBBI atau benar-benar tidak dapat diperbaiki akan dihapus[7].
4. *Filtering:* membuang kata yang tidak memiliki arti atau kata yang sering muncul seperti kata “dari”, “dan”, “yang” dan lain-lain.
5. *Stemming:* mengubah kata berimbuhan menjadi kata dasar sesuai dengan kamus KBBI. System kerja dari stemming adalah menghapus awalan kata(prefiks), akhiran(sufliks) dan gabungan awalan dan akhiran(konflik)[8].

TABEL I
PREPROCESSING TEKS

No	Proses	Hasil
1	Teksasli	TERUNGKAP Penuntut Facebook Sebesar 1 TrivunTernyata ORANG GILA
2	Casefolding	terungkappenuntutfacebooksebesar 1 trivunternyata orang gila
3	Cleansing	terungkappenuntutfacebooksebesartrivunternyata orang gila
4	Correction	terungkappenuntutsebesartriliunternyata orang gila
5	Filtering	terungkappenuntuttriliun orang gila
6	Stemming	ungkaptuntuttriliun orang gila

F. Ekstraksi Fitur

Ekstraksi fitur yang digunakan adalah metode *Term Frequency Term Frequency Inverse Document Frequency* (TF-IDF). Metode TF-IDF adalah metode untuk menghitung bobot setiap kata yang paling umum digunakan pada informasi retrival[9]. Metode TF-IDF merupakan perhitungan *Term frequency* (TF) dengan *Inverse Document Frequency* (IDF) pada tiap term disetiap dokumen dalm korpus. Berikut merupakan rumus dari persamaan TF-IDF.

$$Wdt = TFdt \times IDFt \tag{1}$$

Invers document frequency (IDF) menunjukkanketersediaansebuah kata (term) dalamseluruhdokumen. Semakinsedikitdokumenmengandung term makasemakinbesarnilai IDF.

$$IDFt = \log(\frac{D}{df}) \tag{2}$$

Dimana:

D adalah total dokumen
dfadalahbanyakdokumen yang mengandung term
TF adalahjumlahataufrekuensikemunculan term *t*dalamdokumen*d*
W adalahbobotdokumenke-*d*terhadap kata term ke-*t*

G. Klasifikasi

Tahapan klasifikasi menggunakan algoritme*supportvectormachine*. Algoritma *support vector machine* dibagi menjadi dua yaitu *support vector machine*liniear dan supportvectornon-linear.

1) Support VectorMachine

Support vector machine adalah suatu teknik untuk prediksi, baik dalam kasus klasifikasi maupun regresi. Algoritma ini masuk kelas *supervised learning* dimana dalam implementasinya perlu adanya tahap training menggunakan *sequential training* kemudian disusul tahap *testing*[10].Konsep *support vector machine* (SVM) adalah membentuk *hyperplane* terbaik yang berfungsi untuk memisahkan dua kelas atau lebih. *Hyperplane* merupakan garis atau bidang datar yang memiliki tujuan untuk memisahkan dua kelas data.

$$\textit{Hyperplane} \qquad \vec{w}.\vec{x} + b = 0 \tag{3}$$

Data kelas negative dan kelas positif dapat dirumuskan sebagai berikut:

$$\text{Kelas negatif} \qquad \vec{w}.\vec{x} + b \leq -1 \tag{4}$$
$$\text{Kelas positif} \qquad \vec{w}.\vec{x} + b \geq +1 \tag{5}$$

Persamaan *Support vector machine* sebagai berikut:

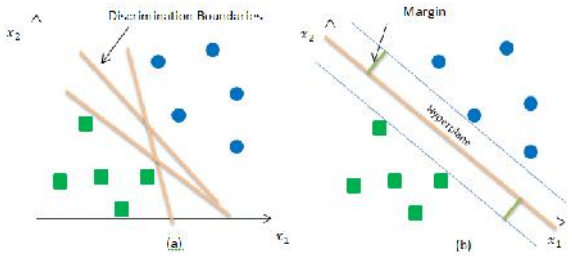
$$f(x) = w.x + b \tag{6}$$

Atau

$$f(x) = \sum_{i=1}^{m} \propto_i y_i K(x_i, x_j) + b \tag{7}$$

Dimana:

w adalah Parameter *Hyperplane* yang dicari
x adalah nilai masukan atribut
b adalah nilai bias

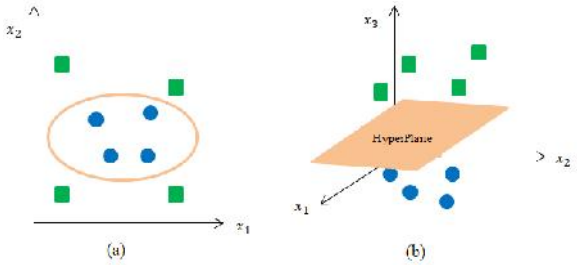


Gambar 6. (A) Gambar *Hyperplane*Yang MungkinDan (B) Gambar *Hyperplane*Dengan Margin Terbaik

2) Support Vector MachineNon-linear&Kernel Trick

Pada dasarnya masalah di dunia nyata (*real world problem*) jarang bersifat *linear* separable dan umumnya masalah bersifat *non-linear*. SVM dimodifikasi dengan memasukan fungsi kernel. Dalam SVM *non-linear*, pertama data (x) akan naikan ruang vector menjadi berdimensi lebih tinggi dengan fungsi ϕ(x)[11].

$$\phi(x_{i,j}) = \phi(x_i) \cdot \phi(x_j) \quad (6)$$



Gambar 7. Fungsi ϕ(X)Memetakan Data Ke Ruang Vector Yang Berdimensi Lebih Tinggi

3) Pengujian

Pengujian dilakukan dua tahap yaitu pengujian pada model dan pengujian pada sampel baru. Pengujian model menggunakan K-*Fold Cross Validation* dan pengujian pada sampel baru menggunakan Confusionmatrix.K-*Fold Cross Validation* adalah teknik validasi dengan membagi data secara acak kedalam K bagian dan masing-masing akan dilakukan proses klasifikasi[12].

Gambar 8. K-Fold Cross Vallidation

Confusion matrix merupakan sebuah metode yang biasa digunakan untuk melakukan pengukuran pada suatu *classifier* dalam melakukan prediksi dari kelas yang berbeda[13].

TABEL II
CONFUSION MATRIX (3X3)

Confusion Matrix		NilaiPrediksi		
		A	**B**	**C**
NilaiSebenarnya	**A**	AA	AB	AC
	B	BA	BB	BC
	C	CA	CB	CC

$$\text{Akurasi} = \frac{AA+BB+CC}{AA+AB+AC+BA+BB+BC+CA+CB+CC} \times 100\% \quad (11)$$

$$\text{Presisi} = \frac{Ai}{Ai+Bi+Ci} \times 100\% \quad (12)$$

$$\text{Recall} = \frac{iA}{iA+iB+iC} \times 100\% \quad (13)$$

III. RESULTS AND DISCUSSION

Hasil penelitian didapatkan dari pengujian akurasi OCR Tesseract, pengujian model SVM *linear* dan *non-linear*, dan pengujian pada data gambar *meme* baru.

A. Pengujian Model SVM Linear Dan Non-Linear

Setelah model selesai dibuat, tahap selanjutnya dilakukan pengujian pada model SVM menggunakan metode K-*Fold Cross Validation*.

TABEL III
HASIL K-FOLD VALIDATION

No	**SVM**			
	Linear	**RBF**	**Polynomial**	**Sigmoid**
1	80.95%	69.05%	69.05%	69.05%
2	76.19%	69.05%	69.05%	69.05%
3	83.33%	78.57%	78.57%	78.57%
4	76.19%	73.81%	73.81%	73.81%
5	73.81%	73.81%	73.81%	73.81%
6	71.43%	71.43%	71.43%	71.43%
7	66.67%	66.67%	66.67%	66.67%
8	71.43%	66.67%	66.67%	66.67%
9	78.57%	71.43%	71.43%	71.43%
10	80.95%	71.43%	71.43%	71.43%
Mean	75.95%	71.19%	71.19%	71.19%

Pengujian akurasi model corpus dilakukan beberapa kategori yang bertujuan untuk mendapatkan model corpus dengan akurasi terbaik. Model corpus terbaik akan digunakan untuk melakukan prediksi pada data uji baru. Berdasarkan Tabel diatas bahwa model SVM *Linear* menggunakan ekstrasi fitur TF-IDF memiliki akurasi terbaik.

B. Pengujian Pada Data Gambar Meme Baru

Setelah didapatkan model dengan akurasi terbaik, selanjutnya mengimplementasikan model untuk menguji gambar *meme* baru. *Meme* yang akan diujikan sebanyak 31 gambar berdasarkan *Control Limit Theorm*. Metode yang digunakan dalam pengujian gambar *meme* baru adalah Confusion Matrix.

TABEL IV
HASIL PENGUJIAN MODEL PADA GAMBAR MEME BARU

Confusion Matrix		**NilaiPrediksi**		
		Negatif	**Netral**	**Positif**
NilaiSebenarnya	**Negatif**	17	1	1
	Netral	2	2	1
	Positif	4	1	2

TABELV
HASIL PENGUJIAN MODEL PADA GAMBAR MEME BARU

Klasifikasi	***Support Vector Machine***	
	Precision	**Recall**
Negatif	74.00%	89.00%
Netral	50.00%	40.00%
Positif	50.00%	29.00%

TABELVI
HASIL AKURASI MODEL PADA GAMBAR MEME BARU

Algoritma	**Ekstraksi Fitur**	**Akurasi**
SVM Liniear	TF-IDF	67.74%

Meme politik dapat dikategorikan kelas positif, jika teks didalam *meme* politik memiliki atribut dengan nilai bobot term yang sesuai dengan nilai bobot term di data training dan *meme* politik memasuki ke ruang kelas positif maka hasil *meme* politik tersebut akan diprediksi sebagai kelas positif. Hal itu juga berlaku pada kelas yang lain.

Suatu *meme* politik yang diprediksi tidak sesuai dengan kelas sebenarnya. Hal ini dikarenakan terdapat beberapa faktor yaitu, Pertama hasil dari OCR Tesseract tidak sesuai dengan teks di dalam gambar *meme* sehingga dalam pembobotan term tidak memiliki nilai bobot sebenarnya. Kedua terdapat kata baru yang belum ada di data training dan tidak memiliki nilai bobot term dan tidak memiliki nilai atribut yang sesuai, sehingga *meme* politik tersebut dapat dekat dengan kelas yang lain dan juga *meme* politik tersebut dapat melewati *Hyperplane* yang dimana dapat memasuki kelas lain.

IV. CONCLUSION

Berdasarkan hasil penelitian yang telah dilakukan bahwa SVM dengan kernel *Linear* memiliki akurasi 75.95%, sedangkan SVM *Non-Linear*(dengan kernel RBF, *Polynomial* dan *Sigmoid*) memiliki akurasi 71.19%. Berdasarkan hasil akurasi setiap kernel dapat disimpulkan bahwa SVM *Linear* memiliki akurasi lebih baik daripada SVM *Non-Linear* dalam melakukan pengklasifikasi *meme* politik. Berdasarkan hasil pengujian data baru *meme*memiliki akurasi rendah yaitu 67.74%. Hal ini terjadi dikarenakan jumlah data *training* negatif lebih banyak daripada data *training* yang lain. Jadi atribut pada data baru lebih condong ke kelas negatif.

Penelitian selanjutnya, diharapakan menggunakan metode lain dalam optical character recognition (OCR), menambahkan fitur object detector yang lain seperti face detector, Menggunakan ekstraksi fitur lain seperti convert negation atau unigram dan menggunakan algoritma klasifikasi lain seperti naïve bayes, decisiontree, K-NN atau neuralnetwork.

ACKNOWLEDGEMENT

TerimakasihkepadaAgiPrasetiadi, S.T., M.Engdan Wahyu Andi Saputra, S.Pd.,M.Engselakudosenpembimbing yang senantiasadalammembimbinglaporanjurnal.

REFERENCES

[1] R. Dawkins, "The Selfish Gene," *Oxford Univ. Press*, 1976.
[2] F. Haisar, "Klasifikasi Analisis Sentimen *Meme* Dengan Metode Optical Character Recognition (OCR) dan Algoritma Naive Bayes," *Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 2, no. 8, 2018.
[3] S. Fanissa, M. A. Fauzi, and S. Adinugroho, "Analisis Sentimen Pariwisata di Kota Malang Menggunakan Metode Naive Bayes dan Seleksi Fitur Query Expansion Ranking," *Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 2, no. 8, pp. 2766–2770, 2018.
[4] F. Jallad, "Object Detection Using Image Processing," *Moscow Inst. Phys. Technol. Dep. Radio Eng. Cybern.*, vol. 1, no. 1, pp. 1–6, 2016.
[5] X. Zhou, C. Yao, H. Wen, Y. Wang, S. Zhou, W. He, and J. Liang, "EAST : An Efficient and Accurate Scene Text Detector," *Comput. Vis. Found.*, vol. 1, no. 1, pp. 1–10, 2015.
[6] V. S. CHANDEL, "Deep Learning based Text Recognition (OCR) using Tesseract and OpenCV," 2018. [Online]. Available: https://www.learnopencv.com/deep-learning-based-text-recognition-ocr-using-tesseract-and-opencv/. [Accessed: 28-Dec-2018].
[7] H. Sujaini and R. D. Nyoto, "Analisis Perbadingan Metode spelling Corrector Peter Norvig dan Spelling Checker BK-Trees pada kata berbahasa indonesia," *J. Sist. dan Teknol. Inf.*, vol. 5, no. 1, pp. 1–5, 2016.
[8] L. Agusta, "Perbadingan Algoritma Stemming Poter dengan Algoritma Nazief-Adriani untuk Stemming Dokumen Teks Bahasa Indonesia," in *Konferensi Nasional Sistem dan Informatika*, 2009, pp. 196–201.
[9] M. Fitri, "perancangan Sistem Temu balik informasi dengan Metode Pembobotan Kombinasi TF-IDF untuk pencarian Dokumen berbahasa indonesia," *Teknol. Inf. dan Komun.*, vol. 1, no. 1, pp. 1–6, 2018.
[10] R. S. Perdana and M. A. Fauzi, "Analisis Sentimen Tingkat Kepuasan Pengguna Penyedia Layanan Telekomunikasi Seluler Indonesia Pada Twitter Dengan Metode *Support vector machine* dan Lexicon Based Features," *Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 1, no. 12, pp. 1725–1732, 2017.
[11] S. R. Kurniasari, "Implementasi SVM dan Asosiasi untuk Sentiment Analysis Data Ulasan The Phoenix Hotel Yogyakarta pada Situs Tripadvisor," 2018.
[12] S. E. Lidya, Syahfitri Kartika, Opim Salim Sitompul, "SENTIMENT ANALYSIS PADA TEKS BAHASA INDONESIA MENGGUNAKAN *SUPPORT VECTOR MACHINE* (SVM)," *Teknol. Inf. dan Komun.*, vol. 2015, no. Sentika, pp. 1–8, 2015.
[13] M. F. Fibrianda and A. Bhawiyuga, "Analisis Perbandingan Akurasi Deteksi Serangan Pada Jaringan Komputer Dengan Metode Naïve Bayes Dan *Support vector machine* (SVM)," *Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 2, no. 9, pp. 3112–3123, 2018.