

**IMPLEMENTASI ALGORITMA *SUPPORT VECTOR MACHINE (SVM)*
UNTUK KLASIFIKASI KONTEN BERITA BAHASA INDONESIA**

SKRIPSI



disusun oleh

Zendi Apri Jeki

15.11.9300

**PROGRAM SARJANA
PROGRAM STUDI INFORMATIKA
FAKULTAS ILMU KOMPUTER
UNIVERSITAS AMIKOM YOGYAKARTA
YOGYAKARTA
2020**

**IMPLEMENTASI ALGORITMA *SUPPORT VECTORMACHINE (SVM)*
UNTUK KLASIFIKASI KONTEN BERITA BAHASA INDONESIA**

SKRIPSI

untuk memenuhi sebagian persyaratan
mencapai gelar Sarjana
pada Program Studi Informatika



disusun oleh

Zendi Apri Jeki

15.11.9300

**PROGRAM SARJANA
PROGRAM STUDI INFORMATIKA
FAKULTAS ILMU KOMPUTER
UNIVERSITAS AMIKOM YOGYAKARTA
YOGYAKARTA
2020**

PERSETUJUAN

SKRIPSI

**IMPLEMENTASI ALGORITMA *SUPPORT VECTOR MACHINE (SVM)*
UNTUK KLASIFIKASI KONTEN BERITA BAHASA INDONESIA**

yang dipersiapkan dan disusun oleh

Zendi Apri Jeki

15.11.9300

telah disetujui oleh Dosen Pembimbing Skripsi
pada tanggal 3 Mei 2019

Dosen Pembimbing,



Mardhiya Hayaty, S.T., M.Kom.

NIK. 190302108

PENGESAHAN

SKRIPSI

IMPLEMENTASI ALGORITMA *SUPPORT VECTOR MACHINE (SVM)* UNTUK KLASIFIKASI KONTEN BERITA BAHASA INDONESIA

yang dipersiapkan dan disusun oleh

Zendi Apri Jeki

15.11.9300

telah dipertahankan di depan Dewan Penguji
pada tanggal 19 Februari 2020

Susunan Dewan Penguji

Nama Penguji

Sri Ngudi Wahyuni, S.T., M.Kom.
NIK. 190302060

Wiwi Widayani, M.Kom.
NIK. 190302272

Mardhiya Havaty, S.T., M.Kom.
NIK. 190302108

Tanda Tangan



Skripsi ini telah diterima sebagai salah satu persyaratan
untuk memperoleh gelar Sarjana Komputer
tanggal 19 Februari 2020



DEKAN FAKULTAS ILMU KOMPUTER

Krisnawati, S.Si, M.T.
NIK. 190302038

PERNYATAAN

Saya yang bertandatangan dibawah ini menyatakan bahwa, skripsi ini merupakan karya saya sendiri (ASLI), dan isi dalam skripsi ini tidak terdapat karya yang pernah diajukan oleh orang lain untuk memperoleh gelar akademis di suatu institusi pendidikan tinggi manapun, dan sepanjang pengetahuan saya juga tidak terdapat karya atau pendapat yang pernah ditulis dan/atau diterbitkan oleh orang lain, kecuali yang secara tertulis diacu dalam naskah ini dan disebutkan dalam daftar pustaka.

Segala sesuatu yang terkait dengan naskah dan karya yang telah dibuat adalah menjadi tanggungjawab saya pribadi.

Yogyakarta, 19 Februari 2020



Zendi Apri Jeki

NIM. 15.11.9300

MOTTO

Khoirunnas anfa'uhum linnas

“Sebaik-baik manusia adalah yang paling bermanfaat bagi manusia”.

(HR. Ahmad, ath-Thabrani, ad-Daruqutni)

Faidza farogh tafanshob

“Maka apabila kamu telah selesai (dari sesuatu urusan), kerjakanlah dengan sungguh-sungguh (urusan) yang lain”.

(Al-Insyirah Ayat 7)

Yarfa'illāhullāzīna āmanū mingkum wallāzīna ūtul-'ilma darajāt

“Allah akan meninggikan orang-orang yang beriman di antaramu dan orang-orang yang diberi ilmu pengetahuan beberapa derajat”.

(Al-Mujadilah Ayat 11)

PERSEMBAHAN

Skripsi berjudul “Implementasi Algoritma *Support Vector Machine (Svm)* Untuk Klasifikasi Konten Berita Bahasa Indonesia” ini dipersembahkan kepada:

1. Allah SWT yang telah memberikan nikmat luar biasa kepada saya.
2. Kedua orang tua saya yang telah memberikan segala yang terbaik kepada anak-anaknya.
3. Universitas AMIKOM Yogyakarta sebagai perguruan tinggi tempat saya bernaung.
4. Ibu Mardhiya Hayaty, S.T., M.Kom. yang telah membimbing saya dalam penyusunan skripsi ini dengan penuh kesabaran.
5. Teman-teman IF-12 angkatan 2015 yang telah menemani saya selama perkuliahan.
6. Para sahabat saya yang tidak bisa saya sebutkan satu per satu.

KATA PENGANTAR

Segala puji bagi Allah SWT tuhan semesta alam yang telah memberikan kenikmatan luar biasa yang tak ada henti-hentinya sehingga skripsi ini dapat terselesaikan dengan sebaik-baiknya. Shalawat serta salam semoga selalu tercurah kepada nabi agung Nabi Muhammad SAW beserta keluarga, para sahabat dan umatnya hingga akhir zaman.

Dengan terselesaikannya skripsi ini yang berjudul “Implementasi Algoritma *Support Vector Machine (SVM)* Untuk Klasifikasi Konten Berita Bahasa Indonesia” penulis mengucapkan terima kasih yang sebanyak-banyaknya kepada:

1. Kedua orang tua penulis bernama Zainal dan Juliana, kakak penulis bernama Zofi Rizki, serta kedua adik penulis bernama Zaldi Yanto dan Ani Yunita yang telah mendukung penuh seluruh kegiatan penulis, sehingga mustahil penulis untuk menghitung kebaikannya.
2. Pasuk Nazirin, S.ST dukungan dari beliau tidak mampu penulis hitung.
3. Bapak Prof. Dr. M. Suyanto selaku Rektor Universitas AMIKOM Yogyakarta.
4. Bapak Sudarmawan, M.Kom. selaku Dekan Fakultas Sains dan Teknologi, dan Ketua Program Studi S1 Informatika.
5. Ibu Mardhiya Hayaty, S.T., M.Kom. selaku pembimbing skripsi ini yang selalu sabar dalam memberikan arahan-arahan serta kesempatan waktu bimbingan yang banyak.

6. Seluruh dosen dan staff Universitas AMIKOM Yogyakarta yang pernah berkontribusi terhadap segala aktivitas penulis selama menjalani perkuliahan.
7. Teman-teman IF-12 angkatan 2015 yang telah menemani penulis selama perkuliahan.
8. Akhi Fail, Cakra Amiyantoro, Aditya Wiha, dan Subraga Islammada yang telah berkontribusi kepada penulis dalam menyelesaikan skripsi ini.
9. Mas Hanif dan Alfi Maghfiroh telah memberikan saran serta masukan.
10. Mas Arif, Ibed, dan Jihad teman seperjuangan mengerjakan skripsi.
11. Teman-teman Sholih di PPM Nur Baiturrahman dan KAMMI Amikom.

Penulis menyadari skripsi ini masih banyak kekurangan, namun penulis berharap skripsi ini dapat memberikan manfaat bagi para pembacanya. Dan semoga Allah SWT membalas kebaikan semua orang yang telah memberikan dukungan dalam bentuk apapun kepada penulis.

Yogyakarta, 19 Februari 2020

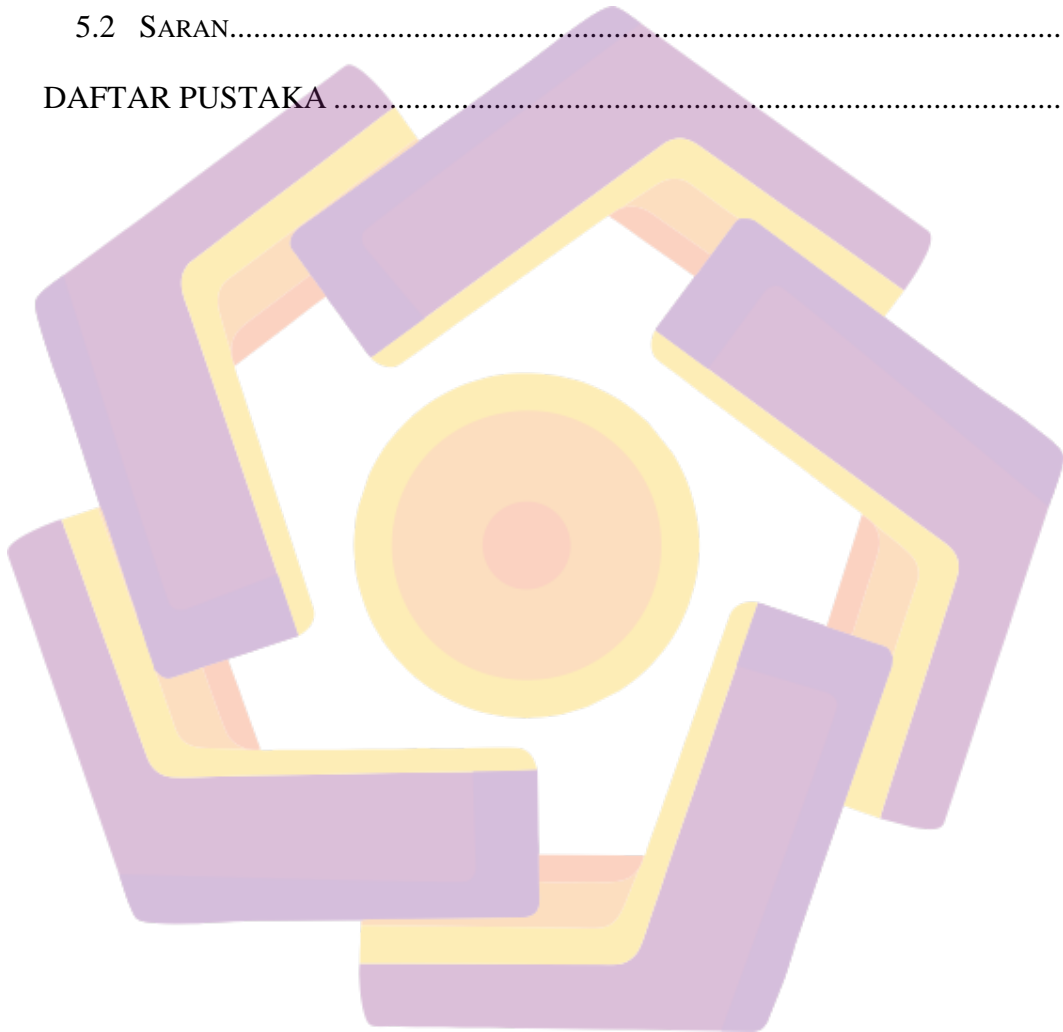
Zendi Apri Jeki

DAFTAR ISI

HALAMAN JUDUL.....	ii
PERSETUJUAN	iii
PENGESAHAN	iv
PERNYATAAN.....	v
MOTTO	vi
PERSEMBAHAN	vii
KATA PENGANTAR	viii
DAFTAR ISI.....	x
DAFTAR TABEL.....	xiii
DAFTAR GAMBAR	xiv
INTISARI.....	xvi
ABSTRACT	xvii
BAB I PENDAHULUAN.....	1
1.1 LATAR BELAKANG.....	1
1.2 RUMUSAN MASALAH	2
1.3 BATASAN MASALAH	2
1.4 MAKSUD DAN TUJUAN PENELITIAN	3
1.5 MANFAAT PENELITIAN.....	3
1.6 METODE PENELITIAN	3
1.7 SISTEMATIKA PENULISAN	4
BAB II LANDASAN TEORI	6
2.1 KAJIAN PUSTAKA.....	6
2.2 LANDASAN TEORI.....	8
BAB III METODE PENELITIAN.....	21
3.1 ALAT PENELITIAN	21

3.2	SUMBER DATASET	21
3.3	ALUR PENELITIAN.....	22
3.3.1	Pengumpulan data.....	24
3.3.2	Preprocessing Data.....	24
3.3.3	Pembagian Data	29
3.3.4	TF-IDF	30
3.3.5	Klasifikasi SVM.....	30
3.3.6	Evaluasi.....	30
3.3.7	Validasi.....	31
3.3.8	Hasil Evaluasi dan Validasi	31
BAB IV HASIL DAN PEMBAHASAN		32
4.1	PENGUMPULAN DATA	32
4.2	PREPROCESSING DATA.....	33
4.2.1	Casefolding	33
4.2.2	Penghapusan Stopwords	34
4.2.3	Cleaning	35
4.2.4	Filtering.....	37
4.2.5	Stemming	37
4.2.6	Tokenizing.....	39
4.3	PEMBAGIAN DATA	40
4.4	PEMBOBOTAN TF-IDF	41
4.5	PELATIHAN DAN PENGUJIAN SVM.....	43
4.6	EVALUASI DAN VALIDASI PERFORMA	45
4.6.1	Evaluasi.....	45
4.6.2	Validasi	46
4.7	HASIL DAN PEMBAHASAN	47
4.7.1	Hasil Evaluasi dengan SVM kernel rbf.....	47
4.7.2	Hasil Evaluasi Dengan SVM kernel linear	48
4.7.3	Hasil Validasi Dengan SVM Kernel rbf	50
4.7.4	Hasil Validasi Dengan SVM Kernel linear	51

4.8 HASIL EVALUASIDAN VALIDASI.....	53
4.8.1 Hasil Evaluasi dengan confusion matrix.....	53
4.8.2 Hasil Validasi dengan K-Fold Cross Validation.....	54
BAB V Penutup	56
5.1 KESIMPULAN.....	56
5.2 SARAN.....	56
DAFTAR PUSTAKA	58



DAFTAR TABEL

Tabel 2.1 Perbedaan Penelitian	7
Tabel 3.1 Spesifikasi Perangkat Keras Dan Lunak	21
Tabel 3.2 Contoh <i>Case Folding</i>	25
Tabel 3.3 Contoh Proses Cleaning	27
Tabel 3.4 Contoh <i>Filtering</i>	27
Tabel 3.5 Contoh <i>Tokenization</i> Menggunakan <i>N-Gram</i>	29
Tabel 4.1 Hasil Evaluasi <i>SVM Kernel Rbf</i>	47
Tabel 4.2 Hasil Evaluasi <i>Kernel Linear</i>	48
Tabel 4.3 Hasil Validasi <i>Kernel Rbf</i>	50
Tabel 4.4 Hasil Validasi <i>Kernel Linear</i>	51

DAFTAR GAMBAR

Gambar 2.1 (A) Pencarian Hyperplane (B) Hyperplane Terbaik	12
Gambar 2.2 Fungsi Φ Memetakan Data Ke Ruang Vector Yang Berdimensi Tinggi	16
Gambar 2.3 Ilustrasi K-Fold Cross Validation	18
Gambar 2.4 Ilustrasi Confusion Matrix	19
Gambar 3.1 Gambaran Umum Alur Penelitian.....	22
Gambar 3.2 Diagram Alur Tahapan Pre-Processing.....	25
Gambar 3.3 Contoh Proses Penghapusan <i>Stopword</i>	26
Gambar 3.4 Contoh Proses <i>Stemming</i> Kata Dengan <i>Sastrawi Stemmer</i>	28
Gambar 4.1 Dataset Format <i>.Csv</i>	32
Gambar 4.2 <i>Import</i> Dataset Menggunakan <i>Library Pandas</i>	33
Gambar 4.3 Implementasi <i>Case Folding</i>	33
Gambar 4.4 Hasil Implementasi <i>Case Folding</i>	34
Gambar 4.5 Sebelum <i>Case Folding</i>	34
Gambar 4.6 Implementasi Penghapusan <i>Stopwords</i>	34
Gambar 4.7 Hasil Implementasi Penghapusan <i>Stopwords</i>	35
Gambar 4.8 Sebelum Penghapusan <i>Stopwords</i>	35
Gambar 4.9 Implementasi <i>Cleaning</i>	36
Gambar 4.10 Hasil Implementasi <i>Cleaning</i>	36
Gambar 4.11 Sebelum <i>Cleaning</i>	36
Gambar 4.12 Implementasi <i>Filtering</i>	37
Gambar 4.13 Hasil Implementasi <i>Filtering</i>	37

Gambar 4.14 Sebelum <i>Filtering</i>	37
Gambar 4.15 Implementasi <i>Stemming</i>	38
Gambar 4.16 Hasil Implementasi <i>Stemming</i>	38
Gambar 4.17 Sebelum <i>Stemming</i>	39
Gambar 4.18 Implementasi <i>Tokenization</i>	39
Gambar 4.19 Hasil Implementasi <i>Tokenization</i>	40
Gambar 4.20 Implementasi Pembagian Data.....	40
Gambar 4.21 Implementasi Tf-Idf	41
Gambar 4.22 Hasil Implementasi Tf-Idf.....	43
Gambar 4.23 Implementasi Pelatihan Dan Pengujian SVM Kernel Rbf.....	44
Gambar 4.24 Implementasi Pelatihan Dan Pengujian SVM Kernel Linear	44
Gambar 4.25 Implementasi Evaluasi Model.....	45
Gambar 4.26 Implementasi Validasi Model	46
Gambar 4.27 Grafik Garis Hasil Akurasi Dengan <i>Confusion Matrix</i>	48
Gambar 4.28 Grafik Garis Hasil Akurasi Dengan <i>Confusion Matrix</i>	49
Gambar 4.29 Grafik Garis Hasil Akurasi <i>K-Fold Cross Validation</i>	52
Gambar 4.30 Grafik Garis Hasil Akurasi <i>K-Fold Cross Validation</i>	53
Gambar 4.31 Grafik Garis Hasil Akurasi <i>Confusion Matrix</i>	53
Gambar 4.32 Grafik Batang Hasil Akurasi <i>Confusion Matrix</i>	54
Gambar 4.33 Grafik Batang Hasil Rata-Rata Akurasi	55
Gambar 4.34 Grafik Garis Hasil Akurasi <i>K-Fold Cross Validation</i>	55
Gambar 4.35 Grafik Batang Hasil Akurasi <i>K-Fold Cross Validation</i>	55
Gambar 4.36 Grafik Batang Hasil Rata-Rata Akurasi	55

INTISARI

Kebutuhan analisis *text mining* sangat diperlukan dalam menangani teks berjumlah besar dan tidak terstruktur (Big data). Salah satu kegiatan penting dalam *text mining* adalah klasifikasi atau kategorisasi teks. Analisis *text mining* dilakukan agar mempermudah kita dalam mengambil informasi atau mengolah informasi yang begitu banyak dari dunia internet atau digital, salah satunya dengan melakukan klasifikasi. Kategorisasi teks memiliki berbagai 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 fungsi kernel.

Pada penelitian ini *support vector machine* akan mengelompokkan berita berdasarkan kategori menjadi 3 bagian atau *class* yaitu : politik, kuliner dan sepakbola. Kernel pada *Support Vector Machine* akan di kombinasikan dengan pembobotan *tf-idf* dan *preprocessing* data.

Dengan trik kernel, metode pembobotan, dan *preprocessing* data diharapkan dapat membantu klasifikasi teks dengan baik serta mampu meningkatkan akurasi. Pengujian dilakukan menggunakan *confusion matrix* klasifikasi dengan metode *support vektor machine* dengan kernel *linear* mendapat rata-rata akurasi 92,7%, kemudian dengan kernel *rbf* mendapat rata-rata akurasi sebesar 92,3%.

Kata Kunci: *tf-idf*, *text mining*, *svm*, *preprocessing*, klasifikasi, konten berita

ABSTRACT

The need for text mining analysis is very much needed in handling large and unstructured text (Big data). One of the important activities in text mining is the classification or categorization of texts. Text mining analysis is carried out to make it easier for us to retrieve information or manage so much information from the internet or digital world, one of them is by doing classification. Text categorization has various approaches including probabilistic approaches, support vector machines, artificial neural networks, or decision tree classification. In learning statistics. Support Vector Machine was chosen because this method has advantages in the field of classification with the help of kernel functions.

In this study support vector machine will group news by category into 3 parts or classes, namely: politics, culinary and football. The kernel in Support Vector Machine will be combined with tf-idf weighting and data preprocessing.

With kernel tricks, weighting methods, and data preprocessing is expected to help text classification well and be able to improve accuracy. Tests carried out using confusion matrix classification with the support vector machine method with linear kernels get an average accuracy of 92.7%, then with kernel rbf get an average accuracy of 92.3%.

Keyword: *tf-idf, text mining, svm, preprocessing, classification, news content*

