

**KLASIFIKASI KATEGORI BERITA
PADA *WEBSITE* JATIMNET.COM BERBASIS *SUPPORT*
VECTOR MACHINE (SVM)**

SKRIPSI



HENDRIYAN OGIVANO

**PROGRAM STUDI S1 SISTEM INFORMASI
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS AIRLANGGA
SURABAYA**

2022

**KLASIFIKASI KATEGORI BERITA
PADA *WEBSITE* JATIMNET.COM BERBASIS *SUPPORT*
VECTOR MACHINE (SVM)**

SKRIPSI



HENDRIYAN OGIVANO

NIM. 081611633005

**PROGRAM STUDI SI SISTEM INFORMASI
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS AIRLANGGA**

SURABAYA

2022

**KLASIFIKASI KATEGORI BERITA
PADA *WEBSITE* JATIMNET.COM BERBASIS *SUPPORT
VECTOR MACHINE (SVM)***

SKRIPSI

**Sebagai Salah Satu Syarat untuk Memperoleh Gelar Sarjana Komputer
Bidang Sistem Informasi pada Fakultas Sains dan Teknologi
Universitas Airlangga**

**Hendriyan Ogivano
NIM. 081611633005**

Disetujui Oleh:

Pembimbing I,

Pembimbing II,



Dr. Rimulio Hendradi, S.Si., M.Si.

NIP. 197102111997021001



Indah Werdiningsih, S.Si., M.Kom.

NIP. 198002172005012001

LEMBAR PENGESAHAN NASKAH SKRIPSI

Judul : Klasifikasi Kategori Berita Pada *Website Jatimnet.com*
Berbasis *Support Vector Machine (SVM)*.
Penyusun : Hendriyan Oglivano
NIM : 081611633005
Pembimbing I : Dr. Rimuljo Hendradi, S.Si., M.Si.
Pembimbing II : Indah Werdiningsih, S.Si., M.Kom.
Tanggal Seminar : 13 Oktober 2022

Disetujui Oleh:

Pembimbing I,

Pembimbing II,


Dr. Rimuljo Hendradi, S.Si., M.Si.
NIP. 197102111997021001


Indah Werdiningsih, S.Si., M.Kom.
NIP. 198002172005012001

Mengetahui:

Ketua Departemen Matematika
Fakultas Sains dan Teknologi
Universitas Airlangga

Koordinator Program Studi Program
Studi S1 Sistem Informasi Universitas
Airlangga


Dr. Herry Suprajitno, S.Si., M.Si.
NIP. 196804041994031020


Dr. Rimuljo Hendradi, S.Si., M.Si.
NIP. 197102111997021001

IR - PERPUSTAKAAN UNIVERSITAS AIRLANGGA

SURAT PERNYATAAN TENTANG ORISINALITAS

Yang bertanda tangan di bawah ini, saya:

Penyusun : Hendriyan Ogivano

NIM : 081611633005

Program Studi : Sistem Informasi

Fakultas : Sains dan Teknologi

Jenjang : Sarjana (S1)

Menyatakan bahwa saya tidak melakukan kegiatan plagiat dalam menulis skripsi saya yang berjudul:

KLASIFIKASI KATEGORI BERITA PADA *WEBSITE* JATIMNET.COM BERBASIS *SUPPORT VECTOR MACHINE* (SVM)

Apabila suatu saat nanti terbukti melakukan kegiatan plagiat, maka saya akan menerima sanksi yang telah ditetapkan.

Demikian surat pernyataan ini saya buat dengan sebenar-benarnya.

Surabaya, 16 Januari 2023

Penulis



Hendriyan Ogivano

iv

SKRIPSI

KLASIFIKASI KATEGORI BERITA...

HENDRIYAN O

iv

SKRIPSI

KLASIFIKASI KATEGORI BERITA...

HENDRIYAN O

PEDOMAN PENGGUNAAN SKRIPSI

Skripsi ini tidak dipublikasikan, namun tersedia di perpustakaan dalam lingkungan Universitas Airlangga, diperkenankan untuk dipakai sebagai referensi kepustakaan, tetapi pengutipan harus seizin penulis dan harus menyebutkan sumbernya sesuai kebiasaan ilmiah.

Dokumen skripsi ini merupakan hak milik Universitas Airlangga

KATA PENGANTAR

Puji syukur penulis panjatkan kehadiran Allah subhanahu wa ta'ala, yang telah melimpahkan rahmat dan karunia-Nya sehingga penulis dapat menyelesaikan penyusunan proposal skripsi yang berjudul “Klasifikasi Kategori Berita Pada *Website Jatimnet.com* Berbasis *Support Vector Machine (SVM)*”.

Penulis tak lupa mengucapkan terima kasih kepada Dr. Rimuljo Hendradi, S.Si., M.Si. selaku dosen pembimbing I dan Indah Werdiningsih, S.Si., M.Kom. selaku dosen pembimbing II, yang dengan sabar dan ikhlas membimbing dan memberikan ilmunya kepada penulis sampai terselesaikannya proposal skripsi ini.

Penulis menyadari bahwa dalam penyusunan skripsi ini banyak terdapat kekurangan, untuk itu penulis senantiasa bersedia dan terbuka dalam menerima saran serta kritik yang bersifat membangun demi kesempurnaan skripsi ini. Semoga skripsi ini dapat memberikan manfaat dan wawasan yang berguna. Aamiin.

Surabaya, 07 September 2022

Penulis,

Hendriyan Ogivano

UCAPAN TERIMA KASIH

Dengan memanjatkan puji syukur kehadiran Allah Subhanahu Wa Ta'ala yang telah melimpahkan segala rahmat, karunia, kesabaran, kekuatan dan hidayahNya, sehingga penyusun dapat menyelesaikan skripsi yang berjudul **“Klasifikasi Kategori Berita Pada Website Jatimnet.com Berbasis Support Vector Machine (SVM)”**. Skripsi ini disusun sebagai upaya salah satu syarat untuk memperoleh gelar Sarjana Komputer (S.Kom.) pada Fakultas Sains dan Teknologi Universitas Airlangga Surabaya.

Dalam pelaksanaan penyusunan skripsi ini memiliki banyak kendala dan mendapatkan bantuan dari banyak pihak, sehingga dengan penuh kerendahan hati, penulis ingin menyampaikan rasa terima kasih yang tulus dan penghargaan yang setinggi-tingginya kepada:

1. Allah SWT dengan segala berkah, rahmat, nikmat, ridho, anugrah dan hidayah-Nya sehingga skripsi ini dapat terselesaikan dengan baik. Beserta Nabi besar Muhammad SAW yang senantiasa memberikan curahan hidayah-Nya.
2. Hendriyan Ogivano. Terima kasih sudah bertahan kuat dan kokoh dalam menyelesaikan fase kehidupan selama masa perkuliahan dan skripsi ini.
3. Dr. Rimuljo Hendradi, S.Si., M.Si. selaku dosen pembimbing I atas pengarahan, kegigihan serta kesabaran dalam membimbing, memberikan ilmu, waktu, motivasi, pengalaman, pelajaran berharga, dan koreksi selama penyusunan skripsi serta memberikan kesempatan kepada penulis untuk menjadi perwakilan mahasiswa UNAIR di Asia University Taiwan sehingga penulis mendapat banyak pengetahuan dan pengalaman yang berhubungan dengan penyelesaian penelitian skripsi ini sehingga penulis dapat menyelesaikan skripsi dengan baik.
4. Indah Werdiningsih, S.Si., M.Kom. selaku dosen pembimbing II atas pengarahan, serta kesabaran dalam membimbing, memberikan waktu dan koreksi selama penyusunan skripsi ini sehingga penulis dapat menyelesaikan skripsi dengan baik.

5. Faried Effendy, S.Si, M.Kom. selaku dosen wali yang telah memberikan bantuan, arahan dan saran selama menjalani masa perkuliahan dengan baik.
6. Seluruh Dosen Sistem Informasi Universitas Airlangga yang telah banyak memberikan ilmu dan fasilitas kepada penulis selama masa studi hingga penulisan skripsi dapat terselesaikan dengan baik.
7. Keluarga di rumah yang selalu mengingatkan untuk mengerjakan skripsi.
8. Soffya Ranti yang selalu mendukung dalam penyusunan skripsi.
9. Affandi, Cendra, Cimeng, dan Haryo yang memberikan saran, solusi, dan tempat dalam pengerjaan skripsi.
10. Keluarga SI'16 KAMI KUAT KAMI BERTEKAD, Man of SI 2016, Aman Enjoy, Pasti Isok, selaku teman-teman penulis yang selalu memberikan suasana kekeluargaan, kerecehan, dan kebersamaan selama masa perkuliahan dan pengerjaan skripsi.

Penyusun menyadari bahwa dalam proposal skripsi ini masih terdapat banyak kekurangan yang memerlukan penyempurnaan. Oleh karena itu masukan berupa kritik dan saran yang membangun sangat diperlukan oleh penyusun untuk dijadikan perbaikan dan penyempurnaan di masa mendatang.

Hendriyan Ogivano, 2022, **Klasifikasi Kategori Berita Pada Website Jatimnet.com Berbasis Support Vector Machine (SVM)**. Skripsi ini di bawah bimbingan Dr. Rimuljo Hendradi, S.Si., M.Si. dan Indah Werdiningsih, S.Si., M.Kom. Departemen Matematika, Fakultas Sains dan Teknologi, Universitas Airlangga, Surabaya.

ABSTRAK

Jatimnet.com adalah platform media online lokal yang berbasis di Surabaya. Jatimnet.com menerbitkan puluhan berita setiap harinya. Pada proses pengiriman berita harus melalui seorang editor yang mana tugasnya mengelompokkan masing-masing berita sesuai dengan kategorinya, hal tersebut memakan banyak waktu karena editor yang bertugas harus memilah terlebih dahulu berita yang sesuai dengan kategori sebelum diterbitkan. Data berita yang diambil pada penelitian ini sebanyak 12.929 berita dengan total 15 kategori yaitu daerah, berita, hukum, politik, nasional, ekbis, olahraga, pendidikan, kesehatan, gaya hidup, tekno, pilkada, infornal, dunia, dan ekonomi. Tujuan penelitian ini adalah mengimplementasikan metode SVM (*Support Vector Machine*) *Multi-class One vs Rest* untuk klasifikasi kategori berita pada website jatimnet.com. Pada penelitian ini menggunakan *library* Scikit-Learn SVM *Linear Kernel*, Scikit-Learn SVM RBF *kernel* dan Spark SVM *Linear kernel*. Pada ketiga variasi metode tersebut akan dibandingkan metode yang memiliki akurasi terbaik dengan cara menggunakan 5 variasi data *split* testing 10%, 20%, 30%, 40%, dan 50% pada setiap metode yang digunakan. Hasil penelitian menunjukkan bahwa akurasi sebesar 80.06% dengan waktu klasifikasi 38 detik untuk Scikit-learn SVM *Linear Kernel* dengan 10% data *split* merupakan akurasi yang terbaik dibandingkan dengan metode lainnya yaitu sebesar 76.34% dengan waktu klasifikasi 19.800 detik untuk Scikit-learn SVM RBF *Kernel* dengan 10% data *split* dan 65,98% dengan waktu klasifikasi 1.200 detik untuk Spark SVM *Linear Kernel* dengan 10% data *split*. Berdasarkan penelitian yang dilakukan klasifikasi kategori jatimnet.com dapat diterapkan dengan menggunakan metode SVM dengan *Multi-class One vs Rest*. Pada skenario proses permodelan klasifikasi pada penelitian ini, dengan memanfaatkan *library* Scikit-learn SVM *Linear Kernel* menghasilkan akurasi yang lebih besar dibandingkan dengan 2 metode lainnya yaitu Scikit-learn SVM RBF *Kernel* dan Spark SVM *Linear Kernel*.

Kata Kunci: Klasifikasi, Berita, SVM, *Support Vector Machine*

Hendriyan Ogivano, 2022, **Classification of News Categories on the Jatimnet.com Website Based on Support Vector Machine (SVM)**. This undergraduate thesis was supervised by Dr. Rimuljo Hendradi, S.Si., M.Si. and Indah Werdiningsih, S.Si., M.Kom. Departemen Matematika, Fakultas Sains dan Teknologi, Universitas Airlangga, Surabaya.

ABSTRACT

Jatimnet.com is a local online media platform based in Surabaya. Jatimnet.com publishes dozens of news every day. The process of sending news must go through an editor whose job is to group each news according to its category, it takes a lot of time because the editor in charge must first sort out the news according to the category before it is published. The news data taken in this research is 12,929 news stories with a total of 15 categories, namely regional, news, legal, political, national, commercial, sports, education, health, lifestyle, techno, pilkada, informal, world, and economics. The purpose of this study is to implement the SVM (Support Vector Machine) Multi-class One vs Rest method for the classification of news categories on the jatimnet.com website. In this study, the Scikit- Learn SVM Linear Kernel, Scikit-Learn SVM RBF kernel, and Spark SVM Linear kernel libraries are used. The three variations of the method will be compared to the method that has the best accuracy by using 5 variations of split testing data 10%, 20%, 30%, 40%, and 50% for each method used. The results show that the accuracy is 80.06% with a time of 38 seconds classification for Scikit-learn SVM Linear Kernel with 10% data split is the best accuracy compared to other methods which are 76.34% with a classification time of 19,800 seconds for Scikit-learn SVM RBF Kernel with 10% data split and 65.98% with Classification time is 1,200 seconds for Spark SVM Linear Kernel with 10% split data. Based on research conducted, Jatimnet.com category classification can be applied using the SVM method with Multi-class One vs Rest. In the scenario of the classification modeling process in this study, using the Scikit-learn SVM Linear Kernel library produces greater accuracy than the other 2 methods, namely Scikit-learn SVM RBF Kernel and Spark SVM Linear Kernel.

Keywords: Classification, News, SVM, Support Vector Machine

DAFTAR ISI

LEMBAR JUDUL	i
LEMBAR PERNYATAAN	ii
LEMBAR PENGESAHAN NASKAH SKRIPSI	iii
SURAT PERNYATAAN TENTANG ORISINALITAS	iv
PEDOMAN PENGGUNAAN SKRIPSI	v
KATA PENGANTAR	vi
UCAPAN TERIMA KASIH	vii
ABSTRAK	ix
ABSTRACT	x
DAFTAR ISI	xi
DAFTAR GAMBAR	xiv
DAFTAR TABEL	xv
DAFTAR LAMPIRAN	xvi
BAB I PENDAHULUAN	1
1.1 Latar Belakang	1
1.2 Rumusan Masalah	3
1.3 Tujuan	3
1.4 Manfaat	3
1.5 Batasan Masalah	3
BAB II TINJAUAN PUSTAKA	5
2.1 Klasifikasi	5
2.2 <i>Application Programming Interface</i> (API)	5
2.3 <i>Support Vector Machine</i> (SVM)	6
2.3.1 <i>Multi-Class Classification</i>	10

2.3.2 <i>Kernel Function</i>	10
2.4 <i>Naive Bayes</i>	12
2.5 <i>Pre-Processing</i>	12
2.5.1 <i>Case Folding</i>	12
2.5.2 <i>Tokenizing</i>	13
2.5.3 <i>Filtering (Stopword)</i>	13
2.5.4 <i>Stemming</i> Nazief dan Adriani.....	13
2.6 <i>Term Frequency-Inverse Document Frequency (TF-IDF)</i>	15
2.7 <i>Python</i>	16
2.8 <i>Scikit-Learn</i>	16
2.9 <i>Confusion Matrix</i>	17
2.10 <i>Jatimnet.com</i>	18
BAB III METODOLOGI PENELITIAN	19
3.1 <i>Objek Penelitian</i>	19
3.2 <i>Tahapan Penelitian</i>	19
3.3 <i>Pengumpulan Data</i>	20
3.4 <i>Pengelompokan Kategori</i>	21
3.5 <i>Proses Pre-processing</i>	23
3.5.1 <i>Case Folding</i>	23
3.5.2 <i>Remove Punctuation</i>	23
3.5.3 <i>Tokenizing</i>	23
3.5.4 <i>Filtering (Stopword)</i>	24
3.5.5 <i>Stemming</i>	24
3.6 <i>Proses Klasifikasi</i>	24
3.6.1 <i>Menghitung Bobot Kata</i>	24

3.6.2 Proses klasifikasi SVM.....	25
3.6.3 Proses klasifikasi <i>Gaussian Naive Bayes</i>	25
3.7 Evaluasi	25
BAB IV HASIL DAN PEMBAHASAN	26
4.1 Pengumpulan Data.....	26
4.2 Pengelompokan Kategori	28
4.3 Pengurangan Data.....	30
4.4 <i>Pre-processing</i>	32
4.4.1 <i>Case Folding</i>	32
4.4.2 <i>Remove Punctuation</i>	32
4.4.3 <i>Tokenizing</i>	34
4.4.4 <i>Filtering (Stopword)</i>	35
4.4.5 <i>Stemming</i>	37
4.5 Proses Klasifikasi.....	38
4.5.1 Menghitung Bobot Kata	38
4.5.2 Proses Klasifikasi dengan SVM	42
4.5.3 Proses Klasifikasi dengan <i>Gaussian Naive Bayes</i>	43
4.6 Evaluasi	44
4.6.1 Evaluasi Model SVM	44
4.6.2 Evaluasi Model <i>Gaussian Naive Bayes</i>	54
4.6.3 Evaluasi dengan Data Terbaru.....	56
4.7 Pembahasan	58
BAB V KESIMPULAN DAN SARAN.....	61
DAFTAR PUSTAKA	62
LAMPIRAN.....	65

DAFTAR GAMBAR

Nomor	Judul Gambar	Halaman
Gambar 2.1	Proses pemanggilan data dari API	6
Gambar 2.2	Ilustrasi <i>support vector machine</i> (SVM) dengan pemisah <i>hyperplane</i> yang optimal pada data linier (Nanda, M. A, et al., 2018)	7
Gambar 2.3	Rumus yang sering digunakan pada metode <i>Support Vector Machine</i> (SVM) (Nanda, M. A, et al., 2018)	11
Gambar 3.1	<i>Flowchart</i> tahap penelitian	19
Gambar 3.2	Persebaran kategori berita	20
Gambar 3.3	<i>Flowchart</i> proses <i>Pre-processing</i>	23
Gambar 4.1	Proses instalasi <i>library</i>	26
Gambar 4.2	Contoh fungsi pengambilan data	27
Gambar 4.3	<i>Error kernel appears to have died.</i>	31
Gambar 4.4	Hasil evaluasi <i>precision</i> pada <i>scikit-learn linear kernel</i>	45
Gambar 4.5	Hasil evaluasi <i>recall</i> pada <i>scikit-learn linear kernel</i>	46
Gambar 4.6	Hasil evaluasi <i>f1-score</i> pada <i>scikit-learn linear kernel</i>	47
Gambar 4.7	Hasil evaluasi <i>precision</i> pada <i>scikit-learn RBF kernel</i>	48
Gambar 4.8	Hasil evaluasi <i>recall</i> pada <i>scikit-learn RBF kernel</i>	49
Gambar 4.9	Hasil evaluasi <i>f1-score</i> pada <i>scikit-learn RBF kernel</i>	50
Gambar 4.10	Hasil evaluasi <i>precision</i> pada <i>Spark linear kernel</i>	51
Gambar 4.11	Hasil evaluasi <i>recall</i> pada <i>Spark linear kernel</i>	52
Gambar 4.12	Hasil evaluasi <i>f1-score</i> pada <i>Spark linear kernel</i>	53
Gambar 4.13	Hasil evaluasi <i>precision</i> pada <i>Gaussian Naive Bayes</i>	54
Gambar 4.14	Hasil evaluasi <i>recall</i> pada <i>Gaussian Naive Bayes</i>	55
Gambar 4.15	Hasil evaluasi <i>f1-score</i> pada <i>Gaussian Naive Bayes</i>	56

DAFTAR TABEL

Nomor	Judul Tabel	Halaman
Tabel 3.1	Persebaran 15 kategori berita	22
Tabel 4.1	Contoh data hasil <i>scraping</i>	28
Tabel 4.2	Penjelasan atribut data	28
Tabel 4.3	Persebaran kategori berita sebelum dipilah	29
Tabel 4.4	Persebaran kategori berita setelah dipilah	30
Tabel 4.5	Persebaran kategori berita setelah pengurangan data	31
Tabel 4.6	Contoh hasil <i>case folding</i>	32
Tabel 4.7	Contoh hasil <i>remove punctuation</i>	34
Tabel 4.8	Contoh hasil dari proses <i>tokenizing</i>	35
Tabel 4.9	Contoh hasil dari <i>filtering (stopword)</i>	36
Tabel 4.10	Contoh hasil <i>stemming</i>	37
Tabel 4.11	Contoh perhitungan <i>Term Frequency</i> (TF)	39
Tabel 4.12	Hasil perhitungan <i>Inverse Document Frequency</i> (IDF)	40
Tabel 4.13	Hasil pembobotan TF-IDF	41
Tabel 4.14	Hasil akurasi menggunakan <i>Scikit-learn linear kernel</i>	42
Tabel 4.15	Hasil akurasi menggunakan Scikit-learn RBF kerne	42
Tabel 4.16	Hasil akurasi menggunakan Spark <i>linear kernel</i>	43
Tabel 4.17	Hasil akurasi menggunakan <i>Gaussian Naive Bayes</i>	44
Tabel 4.18	Persebaran Data Terbaru/ <i>Real</i>	57
Tabel 4.19	Hasil evaluasi <i>Confusion Matrix</i> pada uji coba data terbaru	58
Tabel 4.20	Perbandingan evaluasi SVM <i>Linear Kernel</i> 10% data <i>split testing</i> dengan <i>Gaussian Naive Bayes</i> 20% data <i>split testing</i>	60

DAFTAR LAMPIRAN

Lampiran 1 *Source code scraping data*

Lampiran 2 *Source code pada proses pre-processing*

Lampiran 3 Hasil perhitungan TF-IDF

Lampiran 4 Source code menggunakan Scikit-learn linear kernel

Lampiran 5 Source code menggunakan Scikit-learn RBF kernel

Lampiran 7 *Source code metode Gaussian Naive Bayes*

Lampiran 8 *Source code uji data terbaru/real*