

**PENERAPAN PART-OF-SPEECH FILTERING PADA FEATURE
SELECTION DALAM METODE SUPPORT VECTOR MACHINE
TERHADAP ANALISA SENTIMEN TWITTER MENGENAI
PEMILIHAN GUBERNUR DKI JAKARTA 2017**

SKRIPSI



disusun oleh :

Fregy Damara

13.11.6927

**PROGRAM SARJANA
PROGRAM STUDI INFORMATIKA
FAKULTAS ILMU KOMPUTER
UNIVERSITAS AMIKOM YOGYAKARTA
YOGYAKARTA
2017**

**PENERAPAN PART-OF-SPEECH FILTERING PADA FEATURE
SELECTION DALAM METODE SUPPORT VECTOR MACHINE
TERHADAP ANALISA SENTIMEN TWITTER MENGENAI
PEMILIHAN GUBERNUR DKI JAKARTA 2017**

SKRIPSI

untuk memenuhi sebagian persyaratan
mencapai gelar Sarjana
pada Program Studi Informatika



disusun oleh:

Fregy Damara

13.11.6927

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

PERSETUJUAN

SKRIPSI

**PENERAPAN PART-OF-SPEECH FILTERING PADA FEATURE
SELECTION DALAM METODE SUPPORT VECTOR MACHINE
TERHADAP ANALISA SENTIMEN TWITTER MENGENAI
PEMILIHAN GUBERNUR DKI JAKARTA 2017**

yang dipersiapkan dan disusun oleh

Fregy Damara

13.11.6927

telah disetujui oleh Dosen Pembimbing Skripsi

pada tanggal 25 Agustus 2017

Dosen Pembimbing,



Hartatik, ST, M.Cs

NIK. 190302232

PENGESAHAN

SKRIPSI

**PENERAPAN PART-OF-SPEECH FILTERING PADA FEATURE
SELECTION DALAM METODE SUPPORT VECTOR MACHINE
TERHADAP ANALISA SENTIMEN TWITTER MENGENAI
PEMILIHAN GUBERNUR DKI JAKARTA 2017**

yang dipersiapkan dan disusun oleh

Fregy Damara

13.11.6927

telah dipertahankan di depan Dewan Penguji
pada tanggal 22 Agustus 2017

Susunan Dewan Penguji

Nama Penguji

Tanda Tangan

Yuli Astuti, M.Kom
NIK. 190302146

Bayu Setiaji, M.Kom
NIK. 190302216

Hartatik, ST, M.Cs
NIK. 190302232



Skripsi ini telah diterima sebagai salah satu persyaratan
untuk memperoleh gelar Sarjana Komputer
tanggal 25 Agustus 2017



DEKAN FAKULTAS ILMU KOMPUTER

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

PERNYATAAN KEASLIAN

Saya yang bertanda tangan dibawah ini menyatakan bahwa, Skripsi ini merupakan karya saya sendiri (ASLI), dan isi dalam skripsi ini tidak terdapat karya yang pernah di ajukan oleh orang lain atau kelompok lain untuk memperoleh gelar akademis di suatu Institusi Pendidikan, dan sepanjang pengetahuan saya juga tidak terdapat karya atau pendapat yang pernah ditulis atau diterbitkan oleh orang lain atau kelompok lain, kecuali yang secara tertulis sebagai acuan dalam naskah ini dan disebutkan dalam daftar pustaka.

Segala sesuatu yang terkait dengan naskah dan karya yang telah dibuat adalah menjadi tanggung jawab saya sendiri

Yogyakarta, 25 Agustus 2017



Fregy Damara
13.11.6927

MOTTO



“Aut viam inveniam aut faciam”

I shall either find a way or make one

Robert Sidney, 1588

PERSEMBAHAN

Alhamdulillah segala puji syukur atas kehadiran Allah SWT yang telah memberikan rahmat dan karunia-Nya sehingga karya ini dapat terselesaikan dengan sebaik – baiknya, tidak lepas dari bantuan dan dukungan berbagai pihak.

Skripsi ini saya persembahkan dengan rasa syukur kepada saksi dan penolong seumur hidupku, Allah SWT

Untuk Bapak, Ibu, Adikku *jazakumullah khairan katsiir* atas pengertian dan harapan kalian, semoga Allah kuatkan kami untuk selalu berbakti..

Untuk Bu Hartatik *jazakillah* atas bimbingan, ilmu, alur berfikir, kerja keras dan pengorbanannya..

Semua pihak yang telah banyak membantu dalam penyusunan skripsi ini yang tidak dapat disebutkan satu persatu.

KATA PENGANTAR

Puji syukur penulis persembahkan atas kehadiran Allah SWT yang telah memberikan rahmat dan karunia-Nya sehingga penulis dapat menyelesaikan skripsi dengan judul *“Penerapan Part-Of-Speech Filtering Pada Feature Selection Dalam Metode Support Vector Machine Terhadap Analisa Sentimen Twitter Mengenai Pemilihan Gubernur Dki Jakarta 2017”* dengan sebaik – baiknya. Tidak lupa sholawat serta salam penulis haturkan kepada junjungan umat Nabi Muhammad SAW, yang telah membawa umat Islam dari jaman jahiliyah ke jaman yang penuh ilmu pengetahuan.

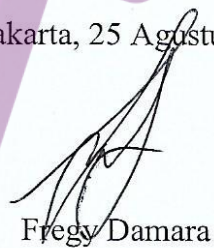
Dengan selesainya skripsi ini, maka penulis mengucapkan terima kasih kepada :

1. Ibu Hartatik, ST, M.CS selaku dosen pembimbing yang telah sabar membimbing dan banyak memberikan pengarahan dan motivasi bagi penulis dalam pembuatan skripsi ini.
2. Bapak Hastari Utama, M.Cs selaku dosen yang membantu dalam penelitian skripsi ini.
3. Wahyu Setiawan, Muhammad Iqbal, Seno Aji, Anggun Putri, Yunisa A, Nabila Nadia, Muhammad Hatta Putra, Afdal Abdullah yang selalu menemani dan turut berpartisipasi dalam pengembangan penelitian ini.
4. Para Dosen dan Staff Universitas Amikom Yogyakarta yang telah banyak memberikan ilmu pengetahuan, pengalaman, dan bantuannya selama penulis kuliah hingga terselesaikannya skripsi ini.

5. Ibu, Bapak, dan Saudara – saudara penulis yang telah memberikan dukungan baik moril ataupun materil.
6. Teman-teman kuliah di kelas TI-03.
7. Teman-teman kontrakan, Arief Darmawan, Nirwan Darmawan, Kurnia Akbar, Fransiskus Paskalis, Izzul Admaza, Joko Budianto, Rizki Pramono, Jecklin Sianturi, Randy Julihartono, Ivan Julio, Ihsan Prasetyo, dan Teguh Kurniawan.
8. Semua pihak yang telah banyak membantu dalam penyusunan skripsi ini yang tidak dapat disebutkan satu persatu.

Penulis tentunya menyadari bahwa pembuatan skripsi ini masih banyak sekali kekurangan – kekurangan dan kelemahan – kelemahannya. Oleh karena itu penulis berharap kepada semua pihak agar dapat menyampaikan kritik dan saran yang membangun untuk menambah kesempurnaan skripsi ini. Semoga skripsi ini dapat bermanfaat bagi pihak terkait dan pembaca pada umumnya.

Yogyakarta, 25 Agustus 2017



Fregy Damara

13.11.6927

DAFTAR ISI

PERSETUJUAN	III
PENGESAHAN	IV
HALAMAN PERNYATAAN KEASLIAN	V
MOTTO	VI
PERSEMBAHAN	VII
KATA PENGANTAR	VIII
DAFTAR ISI	X
DAFTAR GAMBAR	XIV
DAFTAR TABEL	XVII
INTISARI	XIX
ABSTRACT	XX
BAB I PENDAHULUAN	1
1.1 Latar Belakang	1
1.2 Perumusan Masalah	2
1.3 Batasan Masalah	3
1.4 Tujuan Penelitian	3
1.5 Metodologi Penelitian	4
1.6 Sistematika Penulisan	6
BAB II LANDASAN TEORI	7
2.1 Tinjauan Pustaka	7
2.2 Analisa Sentimen	8
2.2.1 Kelas Sentimen	9
2.3 Text Mining	9
2.4 Preprocessing	10
2.5 Part of Speech Tagging (POS Tagging)	11
2.5.1 Stanford POS Tagger	13
2.6 Feature Selection	14
2.6.1 Supervised Feature Selection	14

2.7	Term Frequency-Inverse Document Frequency (TF-IDF)	15
2.8	Cosine Similarity.....	16
2.9	Support Vector Machine	16
2.9.1	Soft Margin	18
2.9.2	Kernel Trick dan non Linear SVM	19
2.9.3	Multi-Class Support Vector Machine	22
2.10	Validasi dan Evaluasi.....	23
2.11	UML (Unified Modeling Language).....	25
2.11.1	Use Case Diagram.....	26
2.11.2	Activity Diagram.....	26
2.11.3	Sequence Diagram	27
2.11.4	Class Diagram	28
BAB III	PERANCANGAN	29
3.1	Analisis Sistem.....	29
3.1.1	Analisis Kebutuhan Fungsional	29
3.1.2	Analisis Kebutuhan Non-Fungsional	29
3.2	Gambaran Umum Sistem	30
3.3	Analisis Algoritma	32
3.3.1	Pengumpulan Data	32
3.3.2	Pelabelan Tweet	33
3.3.3	Pre-processing.....	34
3.3.4	Training.....	46
3.4	Analisis Kebutuhan Data.....	54
3.4.1	Perancangan Database.....	55
3.4.2	Perancangan Dataset	55
3.4.3	Perancangan File Pengumpulan Tweet	56
3.5	Perancangan Sistem	57
3.5.1	Perancangan Use Case Diagram	57
3.5.2	Perancangan Activity Diagram	63
3.5.3	Perancangan Class Diagram.....	67

3.5.4	Perancangan Sequence Diagram	68
3.6	Perancangan Antarmuka Pengguna.....	70
3.6.1	Perancangan Antarmuka Sistem Training.....	70
3.6.2	Perancangan Antarmuka Sistem Prediksi	70
BAB IV	IMPLEMENTASI	75
4.1	Deskripsi Implementasi.....	75
4.2	Implementasi Perancangan Data.....	75
4.2.1	Perancangan Database.....	75
4.3	Implementasi Pengumpulan Data Tweet	76
4.4	Implementasi Pelabelan Data	78
4.5	Implementasi Preprocessing	80
4.5.1	Tweet Cleaning	80
4.5.2	Tokenize	81
4.5.3	Slang Replacement.....	82
4.5.4	POS Tagging	84
4.5.5	POS Filtering.....	85
4.5.6	Stemming	86
4.5.7	Pembentukan Feature List.....	86
4.6	Vektor Tweet.....	88
4.6.1	Implementasi TF-IDF Cosine Similarity	88
4.6.2	Implementasi Pelatihan Support Vector Machine.....	91
4.7	Implementasi Antarmuka	92
4.7.1	Implementasi Antarmuka Pelatihan Model.....	92
4.7.2	Implementasi Antarmuka Prediksi Tweet	95
4.8	Evaluasi Model.....	97
4.8.1	Penerapan POS Filtering	98
4.8.2	Evaluasi Model terhadap Hasil Pelatihan	99
4.8.3	Evaluasi Model terhadap Prediksi.....	102
BAB V	KESIMPULAN.....	106
5.1	Kesimpulan	106

5.2	Saran.....	108
DAFTAR PUSTAKA		109



DAFTAR GAMBAR

Gambar 2.1 (a) Pencarian <i>Hyperplane</i> (b) <i>Hyperplane</i> terbaik [5]	17
Gambar 2.2 Pemetaan <i>input space</i> berdimensi dua dengan pemetaan ke dimensi tinggi [5]	20
Gambar 2.3 Ilustrasi <i>One Against One</i>	23
Gambar 2.4 Ilustrasi <i>One Against All</i>	23
Gambar 2.5 Ilustrasi <i>K-fold cross validation</i>	24
Gambar 2.6 Contoh Perancangan Use Case Diagram	26
Gambar 2.7 Contoh Perancangan Activity Diagram	27
Gambar 2.8 Contoh Perancangan Sequence Diagram	28
Gambar 2.9 Contoh Perancangan Class Diagram	28
Gambar 3.1 Gambaran Umum Sistem	31
Gambar 3.2 Preprocessing	35
Gambar 3.3 Training	46
Gambar 3.4 Perancangan File Training	56
Gambar 3.5 Perancangan File Prediksi	56
Gambar 3.6 Perancangan File Pengumpulan Tweet	57
Gambar 3.7 Activity Diagram Export Model	63
Gambar 3.8 Activity Diagram Import Data Tweet	64
Gambar 3.9 Activity Diagram Prediksi Tweet	64
Gambar 3.10 Activity Diagram Training Data	65
Gambar 3.11 Activity Diagram Upload Model	65
Gambar 3.12 Activity Diagram Upload Tweet	66
Gambar 3.13 Activity Diagram Upload Feature List	66
Gambar 3.14 Class Diagram Training Data	67
Gambar 3.15 Class Diagram Prediksi Data	67
Gambar 3.16 Sequence Diagram Export Model	68
Gambar 3.17 Sequence Diagram Predik Tweet	68
Gambar 3.18 Sequence Diagram Training Data	69

Gambar 3.19 Sequence Diagram Upload Tweet.....	69
Gambar 3.20 Antarmuka Sistem Training	70
Gambar 3.21 Antarmuka Halaman Utama.....	71
Gambar 3.22 Antarmuka Hasil Prediksi	72
Gambar 3.23 Antarmuka Daftar Model	73
Gambar 3.24 Antarmuka Daftar Feature List	74
Gambar 4.1 Command Migrate Server	75
Gambar 4.2 Class Model Training	76
Gambar 4.3 Class Setting	76
Gambar 4.4 Class Testing Data.....	76
Gambar 4.5 Class Feature List	76
Gambar 4.6 Script Pengumpulan Data Tweet.....	77
Gambar 4.7 Hasil JSON Pengumpulan Data	78
Gambar 4.8 Kuisisioner Validasi Data Training	79
Gambar 4.9 Respon Kuisisioner.....	79
Gambar 4.10 Contoh File Data Training.....	80
Gambar 4.11 Tweet Cleaning	81
Gambar 4.12 Script Tokenizer	82
Gambar 4.13 List Slang.....	83
Gambar 4.14 Script Slang Replacement	83
Gambar 4.15 Konfigurasi POS Tagger	84
Gambar 4.16 Script Proses POS Tagging	84
Gambar 4.17 Script Proses POS Filtering.....	85
Gambar 4.18 Script Proses Stemming	86
Gambar 4.19 Script Proses Pembentukan Feature List.....	87
Gambar 4.20 Script Penyimpanan Feature List	88
Gambar 4.21 Script Fungsi Menghitung TF-IDF dan Cosine Similarity	89
Gambar 4.22 Script menghitung bobot per tweet	90
Gambar 4.23 Hasil Script perhitungan bobot per tweet.....	91
Gambar 4.24 Script Konfigurasi Model.....	91

Gambar 4.25 Script Proses Training Data.....	91
Gambar 4.26 Script proses menyimpan model training.....	92
Gambar 4.27 Antarmuka Sistem Pelatihan Model.....	93
Gambar 4.28 Dialog Pemilihan Data Training	94
Gambar 4.29 Pesan Proses Training	95
Gambar 4.30 Halaman Utama Sistem Predik Tweet	96
Gambar 4.31 Halaman Hasil Prediksi Tweet.....	96
Gambar 4.32 Halaman List Model.....	97
Gambar 4.33 List Feature List	97
Gambar 4.34 Script Pembagian Data	98
Gambar 4.35 Grafik Akurasi dari hasil Proses 10-Fold Cross Validation.....	102
Gambar 4.36 Perbandingan Akurasi Pelatihan	103



DAFTAR TABEL

Tabel 2.1 Ilustrasi POS Tagging	12
Tabel 2.2 Tagset	12
Tabel 2.3 Jenis Kernel Trick	21
Tabel 2.4 Confussion Matrix	24
Tabel 3.1 Tweet yang telah di ambil.....	33
Table 3.2 Tweet Berlabel	34
Tabel 3.3 Tweet Cleaning	35
Tabel 3.4 Tokenize.....	37
Tabel 3.5 Normalisasi Kata.....	38
Tabel 3.6 POS Tagging	39
Tabel 3.7 POS Filtering	41
Tabel 3.8 Stemming	43
Tabel 3.9 Feature List	44
Table 3.10 Vektor Feature.....	45
Tabel 3.11 TF Tweet ke 5	47
Tabel 3.12 DF Tweet ke 5.....	47
Tabel 3.13 IDF Tweet ke-5	48
Tabel 3.14 TF-IDF Tweet Ke 5.....	49
Tabel 3.15 Cosine Similarity	50
Tabel 3.16 Nilai W_1, W_2, W_3 , dan b pada setiap label.....	53
Tabel 3.17 Klasifikasi SVM.....	53
Tabel 3.18 Confussion Matrix.....	54
Tabel 3.19 Accuracy, Precision, dan Recall	54
Tabel 3.20 Tabel modelData	55
Tabel 3.21 Tabel Setting	55
Tabel 3.22 TestingData	55
Tabel 3.23 Deskripsi Use Case Training Data	57
Tabel 3.24 Deskripsi Use Case Predik Tweet.....	58

Tabel 3.25 Use Case Import Data Tweet	59
Tabel 3.26 Use Case Export Model	61
Tabel 3.27 Use Case Upload Data Tweet	61
Table 3.28 Use Case Upload Model	62
Tabel 4.1 Hasil Tweet Cleaning	81
Tabel 4.2 Hasil Script Tokenizer	82
Tabel 4.3 Hasil Script Slang Replacement	84
Tabel 4.4 Hasil Script POS Tagging	85
Tabel 4.5 Hasil Script POS Filtering	85
Table 4.6 Hasil Script Stemming	86
Table 4.7 Hasil Script Feature List	87
Tabel 4.8 Perbandingan Jumlah Feature	99
Tabel 4.9 Pengukuran Akurasi, Precisiom, Recall, dan F1-Score	100
Tabel 4.10 Perbandingan Hasil Pelatihan POS Filter dan Tanpa POS Filter	103
Tabel 4.11 Evaluasi Prediksi	104
Tabel 4.12 Confussion Matrix Model POS Filtering	105
Tabel 4.13 Confussion Matrix Model Tanpa POS Filtering	105

INTISARI

Di dalam penelitian ini, peneliti melakukan penelitian untuk mengetahui efek dari feature selection pada Support Vector Machine dalam melakukan klasifikasi sentimen pada tweet di Twitter. Input space yang diberikan terhadap SVM yaitu sebuah feature yang telah diproses melalui tahapan Part-of-Speech Filtering, yang berguna untuk menentukan porsi kata-kata yang sesuai untuk proses pembelajaran model dari perspektif teoritis maupun linguistik. Terdapat 4 tag yang di seleksi yaitu tag kata benda(NN), kata kerja(VB), kata sifat(JJ), dan kata keterangan(RB).

Input Space tersebut sebelumnya telah diolah melalui perhitungan bobot TF-IDF. Setelah TF-IDF masing-masing tweet diketahui, lalu akan diukur kedekatan TF-IDF tersebut dengan feature list positif, negatif, dan netral melalui perhitungan Cosine Similarity. Ketiga bobot Cosine Similarity ini yang nantinya akan di klasifikasi oleh Support Vector Machine.

Kesimpulan yang dihasilkan dari penelitian ini adalah akurasi model yang didapatkan oleh model tanpa proses POS Filtering mengungguli model dengan proses POS Filtering dengan persentase masing-masing 96.66 % dan 99.25 %. Adapun persentase akurasi prediksi yang dilakukan oleh masing-masing model yaitu sebesar 53,33% untuk pos filter dan 56,66% untuk POS Filter.

Kata Kunci : Support Vector Machine, Part Of Speech, Part Of Speech Filtering, Data Mining, Text Mining, Filtering Feature Selection, Analisis Sentimen.

ABSTRACT

In this research, the authors doing a research to determine the effects of feature selection on the Support Vector Machine in classifying sentiments on Twitter tweets. The input space given to SVM is a feature that has been processed through the Part-of-Speech Filtering stage, which is useful for determining the portion of words appropriate for the learning process model from theoretical and linguistic perspectives. There are 4 tags that are selected, that is the noun tag (NN), verb (VB), adjectives (JJ), and adverbs (RB).

Input Space has previously been processed by the calculation of TF-IDF weight. After TF-IDF of each tweet has been calculated, next step is measure the similarity of TF-IDF between each positive, negative and neutral feature list by calculating Cosine Similarity weights. These three Cosine Similarity weights will be classified by the Support Vector Machine.

In addition, the comparison of two models (POS Filter and without POS Filter) clarified that the models without POS Filtering outperformed the model with POS Filtering with the percentage of accuracy by 99,25%. The percentage of accuracy that obtained by the model with POS Filtering is 96.66%. The percentage of prediction accuracy done by each model is equal to 53,33% for filter post and 56,66% for POS Filter. This proves that the number of features in features list used in the Cosine Similarity weighting process has an effect on the classification process done by Support Vector Machine.

Keywords : Support Vector Machine, Part Of Speech, Part Of Speech Filtering, Data Mining, Text Mining, Filtering Feature Selection, Sentiment Analysis.