Mendeteksi Spammers di Twitter dengan SVM Classifier

Tugas Akhir

diajukan untuk memenuhi salah satu syarat

memperoleh gelar sarjana

dari Program Studi Ilmu Komputasi

Fakultas Informatika

Universitas Telkom

1107120128

Damarsasi Cahyo Wilogo

Program Studi Sarjana Ilmu Komputasi

Fakultas Informatika

Universitas Telkom

Bandung

2018

LEMBAR PENGESAHAN

Mendeteksi Spammers di Twitter dengan SVM Classifier

Detecting Spammers on Twitter with SVM Classifier

NIM : 1107120128

Damarsasi Cahyo Wilogo

Tugas akhir ini telah diterima dan disahkan untuk memenuhi sebagian syarat memperoleh gelar pada Program Studi Sarjana Ilmu Komputasi

Fakultas Informatika

Universitas Telkom

Bandung, 26/7/2018

Menyetujui

Pembimbing I,	Pembimbing II,
Erwin Budi S., S.Si., M.T.	Yuliant Sibaroni S.Si., M.T.
00760045	00750036

Ketua Program Studi
Sarjana Ilmu Komputasi,

Dr., Deni Saepudin S.Si., M.Si.
NIP: 99750013

LEMBAR PERNYATAAN

Dengan ini saya, Damarsasi Cahyo Wilogo, menyatakan sesungguhnya bahwa Tugas Akhir saya dengan judul Mendeteksi Spammers di Twitter dengan SVM Classifier beserta dengan seluruh isinya adalah merupakan hasil karya sendiri, dan saya tidak melakukan penjiplakan yang tidak sesuai dengan etika keilmuan yang belaku dalam masyarakat keilmuan. Saya siap menanggung resiko/sanksi yang diberikan jika di kemudian hari ditemukan pelanggaran terhadap etika keilmuan dalam buku TA atau jika ada klaim dari pihak lain terhadap keaslian karya,

Bandung, 26/7/2018

Yang Menyatakan

Damarsasi Cahyo Wilogo

Mendeteksi Spammers di Twitter dengan SVM Classifier

Damarsasi Wilogo[1], Erwin Budi Setiawan, S.Si., M.T. [2], Yuliant Sibaroni, S.Si., M.T.[3]

[1,2,3]Fakultas Informatika, Universitas Telkom, Bandung
[4]Divisi Digital Service PT Telekomunikasi Indonesia
[1]damarsasiwilogo@gmail.com, [2]setiawanerwinbudi@gmail.com, [3]ysibaroni@gmail.com

Abstrak

Dalam Tugas Akhir ini dibahas tentang pemodelan dan simulasi mendeteksi *spammer* di Twitter dengan menggunakan metode *Support Vector Machine* (SVM). Banyaknya spam pada media sosial salah satunya Twitter dapat mempengaruhi pengguna Twitter dalam mendapatkan informasi yang dapat dipertanggungjawabkan kebenaran dari informasi tersebut, sehingga dibutuhkan suatu teknik untuk mendeteksi bahwa suatu konten merupakan spam atau tidak. Maka pada penelitian ini menggunakan metode SVM dalam mengklasifikasi spam. Pemilihan metode SVM ini dikarenakan dari beberapa penelitian bahwa metode ini dapat memberikan hasil yang baik dalam proses klasifikasi. Pada penelitian ini memberikan hasil akurasi sebesar 96.67% pada rasio 90 data training 10 data testing dengan menggunakan seluruh fitur, untuk penggunaan kelompok fitur tweet hasil akurasi tertinggi didapatkan pada rasio 80:20 sebesar 96.67%, dan untuk penggunaan kelompok fitur user hasil akurasi tertinggi didapatkan pada rasio 60:40 sebesar 75%. Dari pengujian tersebut penggunaan kelompok fitur tweet memberikan hasil yang sangat berpengaruh dibandingkan dengan penggunaan kelompok fitur user, hal ini dibuktikan dengan hasil akurasi dari penggunaan kelompok fitur tweet sama dengan hasil akurasi dari penggunaan seluruh fitur.

Kata kunci : Twitter, Support Vector Machine (SVM), Spam, Klasifikasi

Abstract

In this final project discussed about modeling and simulation detecting spammers on Twitter by using Support Vector Machine (SVM) method. Many of spam on social media one of which Twitter can affect Twitter users in getting information that can be justified the truth of the information, so it takes a technique to detect a content is a spam or not, so in this final project using SVM method in classifying spam. The selection of SVM method is because of some research that this method can give good results in the process of classification. In this research, the result of accuracy is 96.67% at 90 for training 10 for testing ratio using all features, for the use of tweet feature group the highest accuracy result is found in 80:20 ratio of 96.67%, and for user feature group usage the highest accuracy result is found in ratio 60:40 by 75%. From these research the use of tweet feature groups gives a very influential result compared to the use of user feature groups, as evidenced by the accuracy of using the tweet feature group equal to the accuracy of the use of all features.

Keywords: Support Vector machine (SVM), Spam, Twitter, Classification

1. Pendahuluan

Latar Belakang

Pada zaman yang sudah modern ini, media sosial merupakan hal yang paling digemari oleh kaum remaja sebagai salah satu cara untuk mengungkapkan apa yang mereka rasakan, ataupun apa yang mereka alami, seperti situasi apa yang terjadi pada lingkungan sekitarnya dan sebagainya. Media sosial sendiri merupakan sebuah media untuk berinteraksi satu sama lain, yang dilakukan secara online di mana hanya pengguna yang terdaftar saja yang dapat melakukan komunikasi.

Salah satu media sosial yang populer saat ini adalah Twitter. Mekanisme Twitter yang membuat tweet tersebar luas, mampu membuat penggunanya membicarakan tentang kejadian, acara, dan mengirim tweet, menjadikan layanan ini memiliki peluang untuk masuknya spam [5,7]. Contohnya, berita *trending topics*, *trending topics* adalah tema dari tweet yang paling banyak dipublikasikan oleh pengguna dalam Twitter. Maka tema itulah yang menjadi sasaran para pelaku spam (*spammer*), mereka membuat tweet berisikan salah satu atau lebih dari tema yang menjadi *trending topics*, *spammer* juga mencantumkan tautan (*link*), dan juga *hashtag*, namun tautan tersebut mengarahkan pada website yang sama sekali tidak berhubungan dengan tema *trending topics* yang dituju, begitu juga dengan *hashtag*, *hashtag* tersebut sama sekali tidak berhubungan dengan tema dari tweet yang dikirim oleh *spammer* [5].

Dalam penelitian sebelumnya [5], untuk mendeteksi *spammer* pada Twitter dengan menggunakan metode SVM dan fitur yang digunakan antara lain *follower*, *following*, *like*, URL, *hashtag*, *spam word*, dan *mention*. Namun, penelitian sebelumnya menggunakan ruang lingkup yang berhubungan dengan tweet dan tidak menggabungkan ruang lingkup tweet dan user. Dalam penelitian tugas akhir ini, penulis masih menggunakan metode SVM. Hal lain yang membedakan penelitian [5] adalah ruang lingkup *spammer* yang berada di Indonesia, menggunakan ruang lingkup yang berhubungan dengan tweet dan user, serta penambahan fitur *follower*, *following*, *spam word Indonesia*, dan *like*.

Pada Tugas Akhir ini topik dan batasannya yaitu bagaimana mengimplementasikan metode SVM dalam mendeteksi *spammer* di Twitter, dan menguji dan menganalisis tingkat akurasi yang diperoleh dari sistem deteksi *spammer* di Twitter dengan menggunakan metode SVM. Batasannya yaitu akun berbahasa Indonesia, data tweet maksimal 100 tweet per akunnya, tidak membandingkan metode SVM dengan metode penelitian lainnya, dan *labeling* kelas dilakukan secara manual.

Tujuan yang ingin dicapai dalam Tugas Akhir ini yaitu mengimplementasikan metode klasifikasi SVM pada deteksi *spammer* di Twitter, menganalisa tingkat nilai akurasi dari sistem deteksi *spammer* di Twitter dengan metode SVM. Data acuan yang digunakan dalam menguji sistem antara lain adalah data user dan data tweet yang diambil pada tanggal 1-6 Juni 2016. Data user terdiri atas beberapa fitur, antara lain *follower*, *following*, *tweet*, dan *like*, serta data tweet terdiri atas URL, *mention*, *hashtag*, *spam words International*, dan *spam words Indonesia*. Terdapat 300 user twitter yang masing-masing datanya sudah terklasifikasi *spammer* dan *non spammer*, dan masing-masing kelas terdiri dari 150 user. Data tersebut diperoleh secara manual dengan mengambil alamat situs resmi twitter.

2. Studi Terkait

2.1 *Crawling Data*

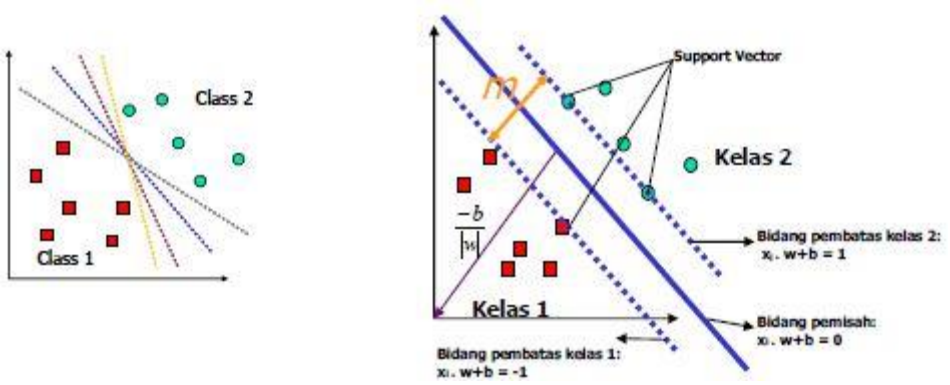
Crawling Data memiliki tujuan untuk mendapatkan data yang akan digunakan sebagai data acuan oleh sistem. Data acuan berupa user dan tweet, data user didapatkan secara manual dari Twitter lalu diberikan label kelas berdasarkan teknik *labeling* yang ada pada penelitian sedangkan untuk data tweet didapatkan melalui Twitter API yang sudah diimplementasikan ke dalam bahasa pemrograman PHP [5]. Data hasil *crawling* kemudian akan dibagi menjadi data *training* dan *data testing*.

2,2 *Pre-Processing*

Tahap pre-processing data bertujuan untuk mendapatkan data acuan yang telah siap diproses ke dalam sistem klasifikasi. Adapun proses yang ada di tahap ini yaitu *Preparation Data*, *Transformation Data*, *Summarization Data*, dan *Merger Data*. *Preparation data* adalah proses membagi data acuan menjadi dua set data yaitu *data testing* dan *data training*. Jumlah rasio pembagian data berdasarkan kebutuhan pengujian sistem. *Data Training* digunakan pada saat proses pelatihan sistem, sedangkan *data testing* digunakan untuk proses validasi sistem yang dibangun. *Data Transformation* adalah proses mengubah fitur yang ada di data mentah ke dalam bentuk fitur yang siap dimasukkan ke dalam klasifikasi, misalnya mengubah fitur *hashtag* yang bernilai 'Yes' menjadi numerik '1'. *Summarization Data* adalah proses mengumpulkan data yang ada di kelompok fitur *tweet*. *Summarization Data* bertujuan untuk menghitung jumlah dari URL, *mention*, *hashtag*, *spam words International*, dan *spam words Indonesia* yang ada di *tweet* setiap *user*. *Data Merger* adalah proses menggabungkan data yang sudah diolah pada *data transformation* dan *data summarization* menjadi satu kesatuan data acuan yang siap diolah ke dalam sistem yang dibangun.

2.3 *Support Vector Machine*

Support Vector Machine (SVM) adalah sistem pembelajaran yang menggunakan ruang hipotesis berupa fungsi-fungsi linier dalam sebuah ruang fitur (*feature space*) berdimensi tinggi, dilatih dengan algoritma pembelajaran yang didasarkan pada teori optimasi dengan mengimplementasikan learning bias yang berasal dari teori pembelajaran statistik [1]. Tujuan dari SVM ini adalah menemukan *hyperplane* terbaik yang memisahkan dua buah kelas pada *input space* [2]. Terdapat dua jenis SVM, pertama, SVM untuk permasalahan *linear* (*linear SVM*) yang dapat memisahkan kelas pada data menggunakan *linear decision boundary* dan yang kedua SVM untuk permasalahan *non-linear* (*non-linear SVM*) yang memisahkan kelas pada data dengan menggunakan *non-linear decision boundary* [3].



Gambar 1 Alternatif bidang pemisah (kiri) dan bidang pemisah terbaik dengan *margin* (m) terbesar (kanan).

Pada **Gambar 1** dapat dilihat berbagai alternatif bidang pemisah yang dapat memisahkan semua data set sesuai dengan kelasnya [4]. Namun, bidang pemisah terbaik tidak hanya dapat memisahkan data tetapi juga memiliki *margin* paling besar. Adapun data yang berada pada bidang pembatas ini disebut *support vector*. Bidang pembatas pertama membatasi kelas pertama sedangkan bidang pembatas kedua membatasi kelas kedua, sehingga diperoleh (2.1):

$$x_i.w + b \geq +1 \text{ untuk } y_i = +1$$
$$x_i.w + b \leq -1 \text{ untuk } y_i = -1 \quad (2.1)$$

Keterangan:
w: normal bidang
b: posisi *bidang relative* terhadap pusat koordinat.

Untuk mengklasifikasikan data yang tidak dapat dipisahkan secara linier, formula SVM harus dimodifikasi karena tidak akan ada solusi yang ditemukan. Oleh karena itu, kedua bidang pembias harus diubah sehingga lebih fleksibel (untuk kondisi tertentu) dengan penambahan *variable* $\xi_i(\xi_i \geq 0, \forall_i : \xi_i = 0$ jika x_i diklasifikasikan dengan benar) menjadi $x_i.w + b \geq 1 - \xi_i$ untuk kelas 1 dan $x_i.w + b \leq -1 + \xi_i$ untuk kelas 2 [4]. Pencarian bidang pemisah terbaik dengan penambahan *variable* ξ_i sering juga disebut *soft margin hyperplane*, sehingga diperoleh (2.2):

$$min\frac{1}{2}|w|^2 + C\left(\sum_{i=1}^{n}\xi_i\right)$$
$$s.t.\, y_i(w.x_i + b) \geq 1 - \xi_i$$
$$\xi_i \geq 0 \quad (2.2)$$

Terdapat beberapa fungsi *kernel* yaitu, *Linear kernel*, *Polynomial kernel*, *Radial Basis Function* (RBF) *kernel*, *Sigmoid kernel*. Fungsi *kernel* yang paling utama dipakai adalah RBF, dikarenakan [6]:

1. *RBF kernel* memetakan sampel *non-linear* ke dalam ruang dimensi yang lebih tinggi dibandingkan *linear kernel*.
2. *RBF kernel* memiliki *hyperparameter* yang lebih sedikit dibandingkan dengan *polynomial kernel*.
3. *RBF kernel* memiliki kesulitan numerik yang lebih sedikit.

2.4 *Performance Measure*

Performance Measure merupakan tahap evaluasi dan analisis dari performa sistem yang kita rancang. Dalam penelitian tugas akhir ini, performa diukur menggunakan *precision*, *recall*, dan nilai akurasi [2]. *Confusion matrix* digunakan sebagai sarana mempermudah perhitungan *precision*, nilai akurasi, dan *recall*. **Tabel 1** merupakan bentuk umum confusion matrix [2].

Tabel 1 *Confusion Matrix*

	Relevant	Not Relevant
Retrieved	True Positive (TP)	True Negative (TN)
Not-Retrieved	False Negative (FN)	False Positive (FP)

True Positive (TP) merupakan nilai kategori hasil klasifikasi dan nilai kategori yang sesungguhnya sama positif. *False Positive* (FP) merupakan nilai hasil kategori hasil klasifikasi positif dan nilai kategori yang sesungguhnya negatif. *True Negative* (TN) merupakan nilai kategori hasil klasifikasi dan nilai kategori yang sesungguhnya sama negatif. Terakhir, *False Negative* (FN) merupakan nilai kategori hasil klasifikasi negatif dan nilai kategori yang sesungguhnya positif [2].

Akurasi merupakan paramater evaluasi terhadap sistem yang dibangun dalam peneiltian tugas akhir ini. Berikut rumusnya [7]:

$$Akurasi = \frac{TP + TN}{(TP + FP + TN + FN)}$$

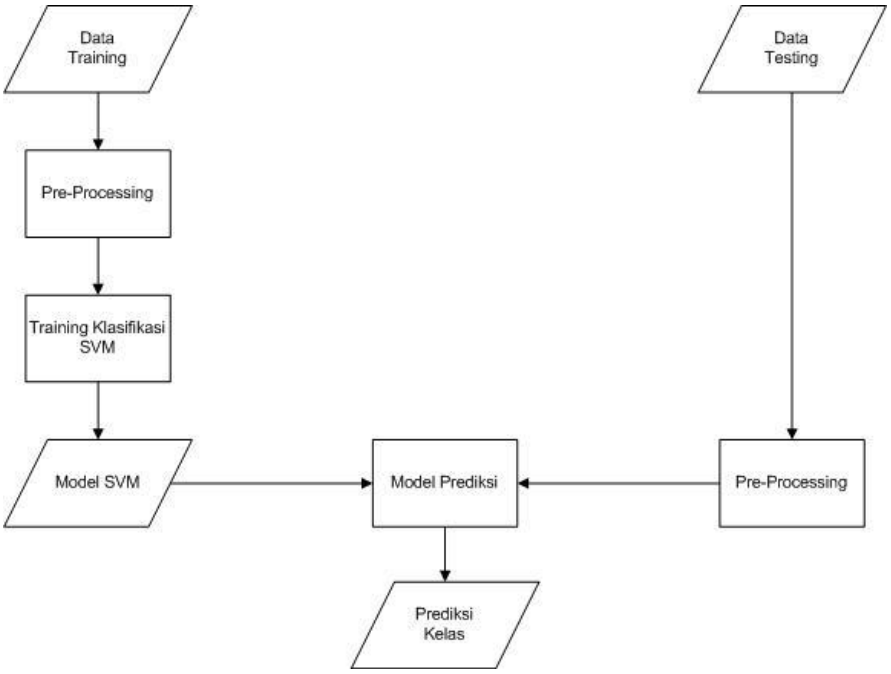
3. Perancangan Desain

Data acuan yang digunakan pada penilitian ini adalah data pengguna Twitter yang diperoleh dengan cara crawling data dari API yang telah disediakan oleh Twitter. Pada proses crawling, fitur menentukan data apa saja yang dibutuhkan, setelah proses crawling, data disimpan ke dalam database untuk selanjutnya dilakukan *pre-processing*. Data Twitter yang digunakan adalah sebanyak 300 user yang terdiri dari 9 fitur, yaitu; *Follower*, *Following*, *Tweet*, *Like*, URL, *Mention*, *Hashtag*, *Spam Word International*, dan *Spam Word Indonesia*, kemudian dilakukan *pre-processing* data meliputi *Preparation Data*, *Transformation Data*, *Summarization Data*, dan *Merger Data*. Hasil dari data *pre-processing* disajikan pada **Tabel 2**.

Tabel 2 Set Data yang Digunakan.

No	Follower	Following	Tweet	Like	URL	Mention	Hashtag	Spam Word International	Spam Word Indonesia	Kelas
1	49747	631	22736	489	22	52	39	0	1	0
2	2575894	481	22500	7	22	80	38	11	14	0
3	7712	2896	44133	20	10	79	62	1	1	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
300	209	925	9509	0	100	19	97	0	0	1

Data acuan terbagi menjadi dua set data yaitu *data training* dan *data testing*. *Data training* digunakan pada proses pembelajaran menggunakan algoritma SVM. Algoritma SVM menghasilkan model *hyperplane* terbaik untuk mengklasifikasi kelas spam atau *non-spam* pada data acuan Twitter. Berikutnya model diuji menggunakan *data testing* untuk mengetahui performansi dari model tersebut. Parameter performansi yang digunakan adalah akurasi, perbandingan data yang terklasifikasi dengan benar dan semua data yang terklasifikasi. Perancangan sistem pada penelitian ini ditunjukkan pada **Gambar 2**.



Gambar 2 Flowchart Pendeteksian Spammer di Twitter.

4. Evaluasi

4.1 Skenario Pengujian

Pengujian pada sistem klasifikasi *spammer* di Twitter bertujuan untuk menganalisis pengaruh jumlah *data training* dan *data testing* dengan metode evaluasi *percentage split* terhadap performansi sistem, menganalisis apakah terdapat gejala *overfitting*, serta menganalisis fitur yang paling berpengaruh di setiap kelasnya. Parameter performansi yang digunakan adalah akurasi. Proses pengujian sistem terdiri dari 2 skenario pengujian.

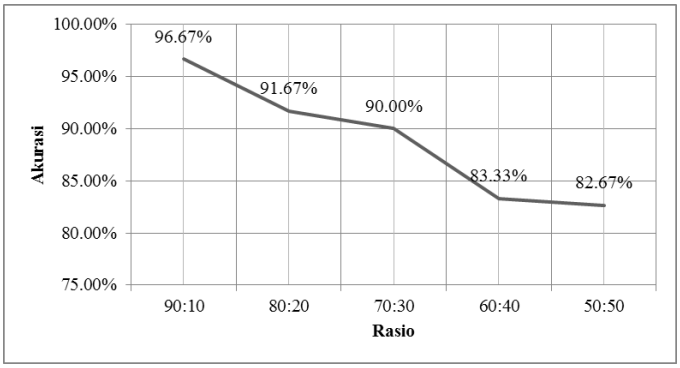
4.1.1 Skenario 1

Pengujian pada skenario 1 menggunakan metode *percentage split*, yaitu metode evaluasi sistem klasifikasi dengan mengubah komposisi *data training* dan *data testing*. Pengujian ini bertujuan untuk menganalisis pengaruh jumlah *data training* dan *data testing* dengan metode evaluasi *percentage split* terhadap performansi sistem, menganalisis apakah terdapat gejala *overfitting* dan menganalisis fitur yang paling berpengaruh di setiap kelasnya. Komposisi *data training* dan *data testing* yang digunakan yaitu dari rasio 90:10 hingga 50:50. Pembagian data pada pengujian skenario ini ditujukan pada **Tabel 3**.

Tabel 3 Skenario Pengujian 1

Rasio	Jumlah Data Training	Jumlah Data Testing
90:10	270	30
80:20	240	60
70:30	210	90
60:40	180	120
50:50	150	150

Sedangkan pengukuran performansi dari pengujian ini menggunakan parameter akurasi disajikan pada **Gambar 3**.



Gambar 3 Hasil Akurasi dari Pengujian Skenario 1.

Pada **Gambar 3** dapat diketahui nilai akurasi paling besar yang didapatkan dari hasil perbandingan antara *data testing* dan *data training* ada pada rasio 90:10 sebesar 96.67%.

4.1.2 Skenario 2

Pengujian pada skenario 2 dilakukan dengan menerapkan metode seleksi fitur dengan membangkitkan kombinasi beberapa fitur lalu memilih kombinasi fitur dengan performansi terbaik. Tujuan dari pengujian ini untuk mencegah gejala *overfitting* yang disebabkan oleh jumlah fitur, serta menganalisis performansi sistem dari *Support Vector Machine* (SVM) menggunakan seleksi fitur terhadap pengujian yang ada di skenario 1.

a) Komposisi Fitur

Pengujian ini dilakukan untuk mengetahui komposisi fitur apa saja yang paling berpengaruh kepada hasil dari akurasi dengan menghilangkan 1 sampai 3 fitur dan terdapat 120 kombinasi komposisi fitur. Pada pengujian ini rasio yang digunakan hanya 90:10.

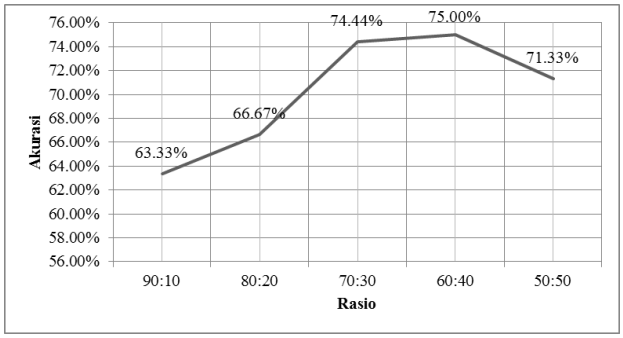
Tabel 4 Akurasi Tertinggi dari Komposisi Fitur.

Komposisi Fitur	Akurasi
3, 5, 6, 7, 8, 9.	96.67%
3, 4, 6, 7, 8, 9.	96.67%
2, 3, 6, 7, 8, 9.	96.67%
2, 3, 5, 6, 7, 9.	96.67%
2, 3, 5, 6, 7, 8.	96.67%
2, 3, 4, 6, 7, 9.	96.67%
2, 3, 4, 5, 6, 7.	96.67%
3, 4, 5, 6, 7, 8, 9.	96.67%
2, 3, 5, 6, 7, 8, 9.	96.67%
2, 3, 4, 6, 7, 8, 9.	96.67%
Keterangan: 1 = Follower, 2 = Following, 3 = Tweet, 4 = Like, 5 = URL, 6 = Mention, 7 = Hashtag, 8 = Spam Word International, 9 = Spam Word Indonesia	

Pada **Tabel 4** menunjukkan bahwa 10 akurasi tertinggi yang didapatkan dari pengujian komposisi fitur adalah sebesar 96.67% dengan didapatkannya nilai akurasi tertinggi dari pengujian komposisi fitur, maka komposisi fitur pada Tabel 4 tersebut berpengaruh untuk menaikkan nilai akurasi dari sistem. Pada pengujian ini dari 120 pengujian dengan 120 kombinasi komposisi fitur didapatkan 18 komposisi fitur yang menghasilkan akurasi yang sama sebesar 96.67%.

b) Hanya Menggunakan Kelompok Fitur User

Fitur user adalah kumpulan dari fitur-fitur yang didapatkan dari segi user. Fitur-fitur yang digunakan adalah *follower*, *following*, *tweet*, dan *like*. Seleksi fitur dengan menggunakan fitur user dilakukan untuk mengetahui fitur yang berpengaruh bagi deteksi *spammer*.

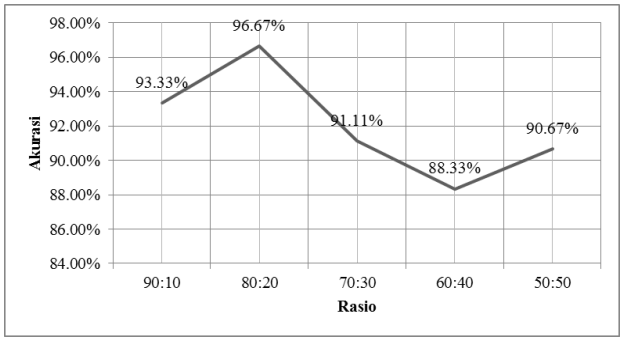


Gambar 4 Hasil Akurasi dari Hanya Menggunakan Kelompok Fitur User.

Pada **Gambar 4** dapat diketahui nilai akurasi paling besar yang didapatkan dari hasil perbandingan antara *data testing* dengan *data training* ada pada rasio 60:40 sebesar 75%.

c) Hanya Menggunakan Kelompok Fitur Tweet

Fitur *tweet* adalah kumpulan dari fitur-fitur yang didapatkan dari segi *tweet*. Fitur-fitur yang digunakan adalah URL, *mention*, *hashtag*, dan *spam words*. Seleksi fitur dengan menggunakan fitur *tweet* dilakukan untuk mengetahui fitur yang berpengaruh bagi deteksi *spammer*.



Gambar 5 Hasil Akurasi dari Hanya Menggunakan Kelompok Fitur Tweet.

Pada **Gambar 5** dapat diketahui nilai akurasi paling besar yang didapatkan dari hasil perbandingan antara *data testing* dan *data training* ada pada rasio 80:20 sebesar 96.67%.

4.2 Analisis Hasil Pengujian

Pada Skenario 1 dan Skenario 2 pengujian dilakukan menggunakan metode *percentage split* dengan mengubah komposisi *data training* dan *data training* sesuai dengan jumlah data yang ada pada Tabel 2. Hasil tertinggi dari pengujian pada Skenario 1 untuk akurasi didapatkan pada rasio 90:10 dengan nilai 96.67%. Berdasarkan hasil dari Skenario 1, didapatlan bahwa semakin tinggi rasio dari data training semakin tinggi pula akurasi yang didapatkan.

Pada Skenario 2 pengujian dilakukan dengan menambahkan penggunaan metode seleksi fitur terhadap pengujian Skenario 1. Pada pengujian komposisi fitur hasil akurasi yang didapatkan dari pengujian komposisi fitur adalah sebesar 96.67% hal ini membuktikan dengan menghilangkan 1 sampai 3 fitur mampu memberikan hasil yang sama dengan menggunakan semua fitur, sehingga semakin sedikit fitur yang digunakan maka kompleksitas dari komputasi semakin rendah. Lalu, berdasarkan hasil 10 komposisi fitur dengan akurasi tertinggi, 3 fitur yang paling relevan dengan *spammer* adalah *Tweet*, *Mention*, dan *Hashtag*. Hal ini disebabkan karena fitur tersebut digunakan paling banyak dari semua hasil kompoisisi dengan akurasi terbaiki. Sedangkan 3 fitur yang tidak relevan dengan *spammer* adalah Like, URL, dan Following.

Hasil tertinggi dari pengujian Hanya Menggunakan Kelompok Fitur User untuk akurasi didapatkan pada rasio 60:40 dengan nilai 75%. Hasil tertinggi dari pengujian Hanya Menggunakan Kelompok Fitur Tweet untuk akurasi didapatkan pada rasio 80:20 dengan nilai 96.67%.

5. Kesimpulan

Berdasarkan hasil pengujian yang dilakukan menggunakan metode SVM, *classifier* dengan rasio data training dan testing 90:10 menghasilkan akurasi terbaik yaitu sebesar 96.67% dibandingkan dengan rasio lainnya. Lalu pada pengujian komposisi fitur, didapatkan 10 dari 120 komposisi fitur yang menghasilkan akurasi terbaik yaitu 96.67% pada rasio 90:10. Dari hasil komposisi fitur dengan akurasi terbaik, didapatkan bahwa fitur *Tweet*, *Mention*, dan *Hashtag* sangat berpengaruh terhadap *spammer*, sedangkan fitur Like, URL, dan Following sangat tidak berpengaruh terhadap kelas *spammer*.

Daftar Pustaka

[1] C. Nello, T. John. 2000. An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. England. Cambridge University.

[2] N. Anto, W. Arief, D. Handoko. 2003. Support Vector Machine Teori dan Aplikasinya dalam Bioinformatika. [Online] Available at: http://asnugroho.net/papers/ikcsvm.pdf [Accessed 11 November 2015].

[3] G. Kumar, G. Ramachandra, K. Nagamani. 2014. An Efficient Feature Selection System to Integrating SVM with Genetic Algorithm for Large Medical Datasets. IJARCSSE.

[4] S. Krisantus. 2007. Tutorial SVM Bahasa Indonesia oleh Krisantus. [Online] Available at: https://www.scribd.com/doc/214404614/Tutorial-Svm-Bahasa-Indonesia-Oleh-Krisantus [Accessed 9 November 2015].

[5] B. Fabricio, M. Gabriel, R. Tiago, A. Virgilio. 2010. Detecting Spammers on Twitter. Brazil. Universidade Federal de Minas Gerais, Belo Horizonte.

[6] V. Gunjan, V. Vineeta. 2012. Role and Application of Genetic Algorithm in Data Mining. International Journal of Computer Applications.

[7] P. David. 2007. Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation. Australia. Flinders University of South Australia.