# Pengaruh Tokenisasi Kata N-Grams Spam SMS Menggunakan *Support Vector Machine*

**Castaka Agus Sugianto [1], Tri Herdiawan Apandi [2]**

*[1] Teknik Informatika , Politeknik TEDC Bandung.*
*[2] Teknik Informatika , Politeknik TEDC Bandung.*
*Jl. Politeknik-Pesantren KM2 Cibabat Cimahi Utara – Cimahi Jawa Barat – Indonesia*
*[1] castaka@poltektedc.ac.id, [2] h.apandi@gmail.com*

**Abstrak**

Pesan singkat atau *Short Message Service* (SMS) merupakan fasilitas yang ada di telepon seluler. Dengan fasilitas pesan singkat ini banyak yang menyalah gunakan pesan singkat tersebut. Spam SMS adalah proses mengirimkan SMS secara terus menerus tanpa di kehendeaki oleh penerimanya baik berupa iklan, jasa maupun penipuan yang dilakukan lewat telepon seluler yang dapat merugikan pengguna. Banyak cara yang yang dilakukan untuk mencegah spam SMS ini. Salah satunya adalah dengan penyaringan spam SMS, penelitian ini bertujuan untuk menyaring spam SMS dengan menggunakan algoritma *Support Vector Machine* (SVM) sehingga dapat memprediksi mana spam SMS dan mana yang bukan spam SMS. Pesan SMS yang dikumpulkan di lakukan preprosessing dengan melakukan kategorisasi, akronim, tokenisasi, *stop word*, pemecahan isi SMS, pembobotan, pemilihan fitur, pembuatan fitur vector, melakukan pelatihan untuk menghasilkan sebuah model. Pada penilitian ini menunjukan hasil akurasi yang terbaik dihasil dari pemecahan kata 4-grams. Sehingga untuk menguraikan/ pemecah kata yang tepat akan menghasilkan akurasi yang baik.

**Kata kunci: SVM, Spam SMS, N-Grams**

***Abstract***

*Short messages or Short Message Service (SMS) is an existing facility on the mobile phone. With this short message facility many who abuse the short message. SMS Spam is the process of sending SMS continuously without being received by the recipient in the form of advertisement, service or fraud done by cellular phone which can harm the user. Many ways are done to prevent SMS spam. One of them is SMS spam filtering, this research aims to filter SMS spam by using Support Vector Machine (SVM) algorithm so as to predict which SMS spam and which are not spam SMS. SMS messages collected by preprosessing by categorization, acronym, tokenisasi, stop word, splitting SMS content, weighting, feature selection, vector feature creation, training to produce a model. In this study showed the best accuracy results from the 4-grams word splitting. So to decipher / solver the right words will produce good accuracy*

***Keywords: SVM, Spam SMS, N-Grams***

## I. PENDAHULUAN

Layananan Pesan singkat atau *Short Message Service* (SMS) merupakan kebutuhan dasar bagi pengguna telepon, SMS adalah pesan pendek berupa teks yang merupakan layanan komponen telepon, web atau sistem komunikasi mobile, yang menerapkan protokol standar komunikasi perangkat telepon seluler. Menurut Asosiasi Telekomunikasi Seluler Indonesia (ATSI) jumlah SMS mencapai 260 miliar SMS yang terkirim pada tahun 2011 dan terdapat 27 ribu *terrabyte* transaksi data (Khemapatapan, 2010).

Penyalahgunaan isi dari SMS yang bisa merugikan penerima sering disebut *Spam* SMS. Fasilitas SMS ini sering disalahgunakan oleh pihak yang tidak bertanggung jawab sering disebut Spam SMS( Anik, 2013). Spam SMS di Amerika Utara, kurang dari 1% dari seluruh sms yang terkirim ditahun 2011, sedangkan pada bagian Asia jumlah spam SMS hingga mencapai 30% yang mengandung spam SMS(Hastie & Tibshirani, 2009). Di karenakan rendahnya hambatan masuk, sehingga banyak *spammers* yang muncul dan jumlah *spam* SMS sangat tinggi. Untuk itu sangat dibutuhkan proses penyaringan SMS terlebih untuk mencegah kerugian yang dihadapi oleh pengguna telepon seluler itu sendiri, *provider*, dan masyarakat umum yang banyak menggunakan layanaan pesan singkat ini.

Pada penelitian sebelumnya mengenai penyaringan spam SMS banyak metode yang penggabungan *Support Vector Machine* dan token memiliki kekurangan pada saat isi dari SMS menggunakan imbuhan jadi tidak maksimal (Hastie & Tibshirani, 2009). *NaiveBayes*("SMS Spam Overview," 2012), algoritma ini memberikan hasil bahwa waktu pemrosesannya lebih cepat dan tinggkat akurasi yang wajar serta waktu belajar dari algoritma *NaiveBayes* lebih cepat dibandingkan dengan algoritma *Decision Trees*. Kemudian *sampling algorithm*(Hu & Yan, 2010). *Neural Network* memberikan *error* generalisasi yang lebih besar daripada *SVM (Almeida et al., 2011)*. Proses pemecah kalimat menjadi kata dan banyaknya jumlah fitur mempengaruhi nilai akurasi yang dihasilkan, pada penelitian ini pemecahan kata hasil berdasarkan pemilihan 1 kata saja. Sehingga kata "selamat" pada spam SMS dan "selamat" ham SMS mempunyai bobot yang sama, ini akan membingungkan *machine learning* (Apandi & Sugianto, 2015). Untuk itu

pemecahan kata menggunakan N-grmas diperlukan untuk membedakan kata yang sama dengan makna yang berbeda, N-Gram dapat diartikan berfungsi dalam pengambilan potongan n karakter dalam suatu string atau kalimat tertentu(Daniel, 2016).

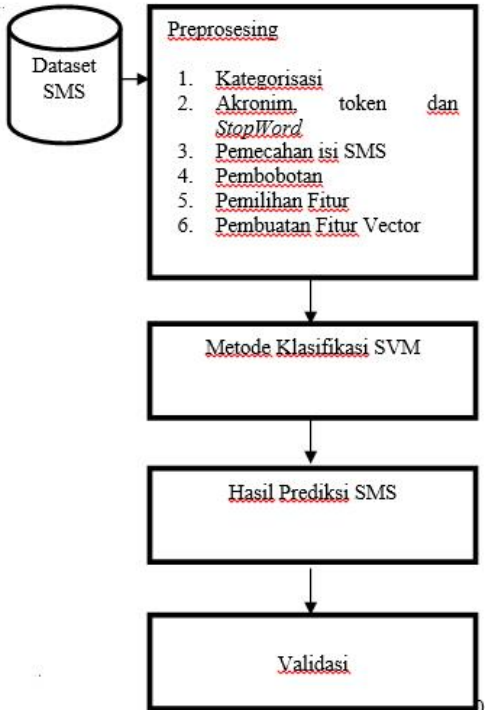
Proses training pada algoritma *Support Vector Machine (SVM)* dilakukan untuk menentukan posisi optimal dari *hyperplane* di *dual space* sedangkan hal ini margin digunakan untuk memisahkan jarak antara fungsi pemisah (*separating hyperplane*) ke masing-masing kelas. Diharapkan pada saat proses taining SVM mencari untuk menentukan *training* set yang paling sempurna. Sedangkan pada metoda yang lain dalam proses training untuk mendapatkan kedua kelas secara optimal atau juga *local* optimal perlu dilakukan secara berulang sampai menemukan *local* optimal (Durgesh, 2009). Ini tentu berbeda dengan SVM *training* dilakukan hanya sekali dengan mendapatkan nilai yang optimal, hal ini memcegah terjadinya *overfitting* karena *overtrained*. Berdasarkan hasil pada penelitian sebelumnya, dapat disimpulkan bahwa algoritma SVM memiliki kinerja dasar terbaik (Almeida et al., 2011). Penelitian ini bertujuan untuk mengetahui pengaruh hasil akurasi tokenisasi kata N-Gram dengan menggunakan SVM.

## II. METODE PENELITIAN

### 1. Tipe Metode Penelitian

Dalam penelitian kali ini menggunakan metode penelitian eksperimen, karena responden berkelompok berdasarkan beberapa kriteria sesuai dengan yang ditugaskan , dengan kata lain juga disebut perlakuan kondisi atau perlakuan *variable* (Sugianto, 2015).

### 2. Metode yang diusulkan




Gambar 1. Metode yang diusulkan

Permodelan yang disajikan pada gambar 1 dimulai dari pengumpulan dataset sampai validasi dataset. Pada fase pertama ini dataset akan dilakukan kategorisasi, akronim, token, *stop word*, pemecahan isi sms, pembobotan, pemilihan fitur, pembuatan fitur *vector*, pada tahap berikutnya hasil dari pengkategorian akan diklasifikasikan dengan metode *Support Vector Machine*, dan fase terakhir akan divalidasi. Penjelasan gambar 1 adalah sebagai berikut.

a. Dataset : dataset yang dikumpulkan adalah 900, baik berupa spam sms dan yang bukan spam. Rincian dari data spam adalah 450 dan data yang bukan spam/ham adalah 450. Dataset ini akan dibagi menjadi 2, untuk training dan testing. Training sebanyak 800 data dan testing 100 data.

b. Kategori text : menentukan mana sms yang spam dan yang bukan/ham. Dengan memberi label spam non spam. Cara untuk membedakan spam dan bukan adalah dengan cara memberikan label dari tiap- tiap data. Data yang spam diberi label 1 dan yang bukan/ham diberi 0.

c. Akromin : mengganti singkatan dengan kata bakunya. Contoh dari tahap ini adalah “jt” menjadi “juta”, “bb” menjadi “blackberry”, dan “dg” menjadi “dengan sesuai”.

Tabel 1. Proses *stopwords*

| Sebelum stop word | Setelah stop word |
|---|---|
| bapak saya mau beli sp three yang 2 gb bisa cod an kapan y? dimana? | bapak saya mau beli sp three2 gb bisa cod an kapan y? dimana? |
| penawaran istimewa...!!! instan cash credit 100juta=3.056.677/bulan legal proses syarat fotocopy : ktp, kartu kredit hubungi andre : 081382162051. | penawaran istimewa...!!! instan cash credit 100juta.056.677/bulan legal proses syarat fotokopi : ktp, kartu kredit hubungi andre : 081382162051. |

d. Stopwords : membuang kata –kata sambung yang sering muncul dan tidak bermakna apa-apa. Maksud tidak bermakna apa-apa contohnya dapat dilihat di tabel 1.

e. Tokenisasi : membuang karakter yang tidak perlu, seperti !@#$%^&*()

f. Pemecahan isi SMS : pada tahap ini menggunakan algoritma N-Grams ini untuk membantu dalam mengambil potongan-potongan kata dari suatu kalimat SMS. Pada metode yang diusulkan menggunakan 3-grams, 4-grams, dan 5 grams kata.

g. Pembobotan : Tahap ini dilakukan pembobotan mengunakan teknik pembobotan Biner, *Document Frequency* (DF) dan *Term Frequency* (TF).

h. Pemilihan Fitur : pemilihan jumlah fitur berdasarkan banyaknya jumlah fitur yang muncul. Jumlah fitur yang digunakan adalah fitur yang mempunyai rangking 50, 10, 150 dan yang terakhir 200.

i. Pembuatan Fitur Vector : mengganti urutan dari bobot yang sudah dipilh menjadi nilai bobotnya, terlebih dulu diberi tanda 1 untuk spam dan 0 untuk ham sms. Contoh pembentukan sebuah kalimat:

Kejutan selamat kepada no simcard mendapatkan hadiah uang tunai rp juta dari pt mtronik pin juta untuk info klik www ptmtronik jimdo com cs

Setelah dilakukan proses pembentukan vektor fitur akan menjadi seperti dibawah ini:

-1 1:2 4:1 6:1 8:1 10:1 11:1 17:1 21:1 26:1 29:1 33:1 34:1 37:1 38:1 39:1 40:1 41:1 42:1 50:1 51:1 53:1

j. Metoda klasifikasi : metode yang digunakan adalah *Support Vector Machine* (SVM) untuk klasifikasi. Pada klasifikasi menggunkan SVM ini akan menghasilkan model yang dipakai untuk menguji data testing.

k. Hasil prediksi : hasil prediksi dari data training berupa akurasi dari tiap-tiap model.

l. Validasi : Pengujian dilakukan menggunakan k-flip lintas validasi teknik. *Cross-validasi* metode digunakan untuk memprediksi keakuratan data pengujian (Hastie & Tibshirani, 2009).

## III. HASIL DAN PEMBAHASAN

Data yang dikumpulkan sebanyak 900 data SMS baik yang data SMS spam maupun data yang bukan spam. Setelah terkumpul data dilakukan proses pengolahan tahap awal yaitu pemberian label, akronim, token dan *stop word*. Berikutnya dilakukan proses pemecahan kalimat menjadi kata, isi dari SMS dipecah – pecah kedalam N-Gram, kemudian dilakukan proses pembobotan dengan teknik pembobotan Biner, *Document Frequency* (DF) dan *Term Frequency* (TF). Kemudian dilakukan kelasivikasi, setelah itu akan dilihat model mana yang memiliki akurasi yang tinggi dari hasil proses pengujian yang dilakukan.

### 1. Pengujian

Dalam pengujian ini dilakukan pemilihan jumlah fitur dari 50 sampai dengan 200 fitur dengan pemecah kalimat kedalam kata 3-grams, 4-grams, 5-grams dengan menerapakan teknik pembobotan biner, *document frequency* dan *term frequency*. Adapun sekenario pengujian bisa dilihat pada tabel 2.

Tabel 2. Sekenario Pengujian Model

| Nama Model | Jumlah Fitur | Pemecah Kalimat | Teknik Pembobotan |
|---|---|---|---|
| Model 1 | 50 | 3-grams kata | Biner |
| Model 2 | 100 | | |
| Model 3 | 150 | | |
| Model 4 | 200 | | |
| Model 5 | 50 | | Document Frequency |
| Model 6 | 100 | | |
| Model 7 | 150 | | |
| Model 8 | 200 | | |
| Model 9 | 50 | | Term Frequency |
| Model 10 | 100 | | |
| Model 11 | 150 | | |
| Model 12 | 200 | | |
| Model 13 | 50 | 4-grams kata | Biner |
| Model 14 | 100 | | |
| Model 15 | 150 | | |
| Model 16 | 200 | | |
| Model 17 | 50 | | Document Frequency |
| Model 18 | 100 | | |
| Model 19 | 150 | | |
| Model 20 | 200 | | |
| Model 21 | 50 | | Term Frequency |
| Model 22 | 100 | | |
| Model 23 | 150 | | |
| Model 24 | 200 | | |
| Model 25 | 50 | 5-grams kata | Biner |
| Model 26 | 100 | | |
| Model 27 | 150 | | |
| Model 28 | 200 | | |

| Nama Model | Jumlah Fitur | Pemecah Kalimat | Teknik Pembobotan |
|---|---|---|---|
| Model 29 | 50 | | Document Frequency |
| Model 30 | 100 | | |
| Model 31 | 150 | | |
| Model 32 | 200 | | |
| Model 33 | 50 | | Term Frequency |
| Model 34 | 100 | | |
| Model 35 | 150 | | |
| Model 36 | 200 | | |

Pada table 2. adalah sekenario pengujian yang dilakukan, dimana terdapat 36 model yang digunakan. Setiap model mempunyai jumlah fitur dan pemecah kalimat yang berbeda-beda. Sekenario pengujian model ini akan menjadi dasar perbandingan.

Pada sekenario table 2 pada dilihat jumlah fitur yang diambil 50 sampai dengan 200. Fitur ini adalah hasil dari pemecahan kata yang teratas dipilihlah 50, 100, 150 dan 200 fitur yang teratas tersebut.

Pada sekenario table 2 terdapat rincian sebagai berikut model 1 sampai dengan 12 menggunakan 3-grams kata, 13 sampai 24 menggunakan 4-grams kata dan sisanya menggunakan 5-grams kata, dengan menggunakan banyak model pemecahan kalimat akan memperkaya hasil perbandingannya.

Pada sekenario table 2. Menggunakan jumlah fitur yang bervariasi dari 50 sampai dengan 200 fitur. Fitur tersebut dipilih karena fitur tersebut yang sering muncul pada dataset.

Terakhir adalah penggunaan teknik pembobotan. Teknik yang dipakai adalah biner, document frequency dan term frequency.

Selanjutnya dari sekenario pada table 1 maka akan dilakukan pengujian akurasi. Pengujian yang dilakukan dengan terlebih dahulu membuat model yang akan diuji tingkat akurasinya. Hasil akurasi dapat diperlihatkan pada tabel 3.

Tabel 3. Hasil Akurasi Dari Model

| Nama Model | Akurasi Pengujian |
|---|---|
| Model 1 | 91% |
| Model 2 | 91% |
| Model 3 | 91% |
| Model 4 | 89% |
| Model 5 | 90% |
| Model 6 | 54% |
| Model 7 | 54% |
| Model 8 | 90% |
| Model 9 | 90% |
| Model 10 | 89% |
| Model 11 | 54% |
| Model 12 | 54% |
| Model 13 | 92% |
| Model 14 | 91% |
| Model 15 | 90% |
| Model 16 | 86% |
| Model 17 | 90% |
| Model 18 | 60% |
| Model 19 | 61% |
| Model 20 | 53% |
| Model 21 | 57% |
| Model 22 | 60% |
| Model 23 | 61% |
| Model 24 | 61% |
| Model 25 | 87% |
| Model 26 | 50% |
| Model 27 | 82% |
| Model 28 | 78% |
| Model 29 | 90% |
| Model 30 | 62% |
| Model 31 | 70% |
| Model 32 | 70% |
| Model 33 | 90% |
| Model 34 | 62% |
| Model 35 | 68% |
| Model 36 | 69% |

Setelah dilakukan uji coba pada datasetnya maka diperoleh hasil akurasi yang diperlihatkan pada tabel 3.

Pada tabel 3 diperoleh akurasi yang tertinggi pada model 13, model 13 adalah pemecahan kata 4-*grams* kata dan menggunakan pembobotan biner. Ini menunjukan dataset yang dikumpulkan rata-rata memiliki kesamaan 4-grams kata, sehinggan mempengaruhi tingkat akurasinya.

Akurasi terendah ditunjukan pada model 26 menggunakan 5-*grams* kata dan menggunakan biner. Sedangkan menggunakan 5-*grams* kata memiliki tingkat akurasi yang

terendah dibandingkan yang lain. Ini menunjukan dengan menggunakan 5-*grams* kata hanya ada dibeberapa SMS saja, tidak merata diseluruh dataset.

Pada pemilihan fitur dengan banyak 50 fitur menunjukan hasil akurasi yang stabil disemua tipe pembobotan baik dari *biner*, *document frequency* dan *term frequency*. Pemilihan fitur 50 teratas menghasilkan akurasi yang stabil menunjukan bahwa semakin banyak fitur yang dipilih semakin tidak konsisten hasil akurasinya, ini disebabkan banyak nilai bobot fitur 100, 150 dan 200 memiliki nilai bobot 1. Nilai bobot satu menunjukan fitur itu hasil sedikit dikumpulan dataset.

Dilihat dari data pengujian akurasi menunjukan bahwa pembobotan menggunakan *term frequency* dan *biner* menunjukan hasil akurasi yang stabil, dikarenakan pembobotan menggunakan *term frequency* dan *biner* akan melihat seluruh data fitur yang berada di dataset. Berbeda dengan *document frequency* menjadikan pembobotan ini memiliki hasul akurasi yang tidak stabil, dikarenakan fitur yang dipilih hanya hasil yang berada di dalam satu *document*/SMS bukan diseluruh dataset.

## IV. KESIMPULAN

Pada penelitian ini menunjukan bahwa pemilihan 50 fitur dan menggunakan 4-grams karakter menunjukan hasil akurasi yang sangat baik dan stabil baik dengan dengan pemilihan fitur *document frequency*, *term frecuency* dan *biner*. Pemilihan jumlah pemecahan kata sangat berpengaruh pada hasil akurasi. Pengunaan pembobotan *document frequency* dan *biner* menunjukan hasil yang baik tetapi tidak terlalu signifikan peneliti menyarankan menggunakan pembobotan *matual information* untuk melihat keterhubungan antara 1 fitur dengan fitur lain pada penelitian selanjutknya.

## DAFTAR PUSTAKA

Almeida, T. A., Gomes, J. M., & Yamakami, A. (2011). Contributions to the Study of SMS Spam Filtering : New Collection and Results, 1–4.

Anik, M,(2013). Collaborative Filtering SMS Spam Berbahasa Indonesia Menggunakan Algoritma Naïve Bayes

Apandi, T. H., & Sugianto, C. A. (2015). Penyaringan Spam Short Message Service Menggunakan Support Vector Machine. In *Seminar Nasional Teknologi Informasi dan Komunikasi Terapan (SEMANTIK)* (pp. 111–116).

Daniel Jurafsky & James Martin, Speech and Language Processing Second Edition. New Jersey: Pearson Prentice Hall,2016.

Durgesh K. Srivastava & Lekha Bhambhu (2009)DATA CLASSIFICATION USING SUPPORT VECTOR MACHINE. in Journal of Theoretical and Applied Information Technology.

Hastie, & Tibshirani. (2009). Cross Validation Bootstrap Methods, 18–26.

Hu, X. I. A., & Yan, F. U. (2010). SAMPLING OF MASS SMS FILTERING ALGORITHM BASED ON FREQUENT TIME-DOMAIN AREA, 548–551. https://doi.org/10.1109/WKDD.2010.50

Khemapatapan, C. (2010). Thai-English Spam SMS Filtering, 226–230.

SMS Spam Overview. (2012). *Cloudmark*, pp. 1–7.

Sugianto, C. A. (2015). ANALISIS KOMPARASI ALGORITMA KLASIFIKASI UNTUK MENANGANI DATA TIDAK SEIMBANG PADA DATA KEBAKARAN HUTAN. *Techno.com*, *14*(4), 336–342.