IJCIT (Indonesian Journal on Computer and Information Technology)
Vol.2 No.2, November 2017, pp. 7~13
ISSN: 2527-449X
E-ISSN**:** 2549-7421

Komparasi Algoritma Support Vector Machine, Naïve Bayes Dan C4.5 Untuk Klasifikasi SMS

Retno Sari
STMIK Nusa Mandiri
e-mail: bee.retno@gmail.com

Abstrak

Layanan pesan singkat atau yang dikenal sebagai SMS, merupakan salah satu cara untuk berkomunikasi oleh para pengguna telepon genggam. SMS terdapat dua macam yaitu sms spam dan sms ham. SMS yang masuk kedalam kotak pesan banyak mengandung SMS yang merupakan spam. Komparasi algoritma Support Vector Machine, Naïves Bayes dan C4.5 untuk klasifikasi sms ini, untuk mengetahui algoritma mana yang memiliki tingkat akurasi yang tinggi. Dapat dilihat dari hasil pengklasifikasian dengan menggunakan 3 metode aplikasi data mining untuk sms berbahasa Indonesia dengan jumlah data sms 200. Akurasi yang didapat dengan menggunakan Naïves Bayes yaitu sebesar 95.00%, sedangkan yang menggunakan Support Vector Machine sebesar 76.00% dan dengan C4.5 akurasi didapat sebesar 95.50%.
Katakunci: *Klasifikasi SMS, Support vector Machine, Naïves Bayes, C4.5*

Abstract

Short message service or known as a text message , is one way to communicate by users mobile phone. SMS there are two kinds of the sms junk and sms human rights. The sms in to the inbox message contain many the sms is junk. Comparation algorithm Support Vector Machine, Naïves Bayes and C4.5 for the classification of sms, to know algorithm which is highly accurate. Can be seen from classification with use 3 method data mining aplication for sms Indonesian with the amount of data sms 200. Accuracy obtained by using Naïves Bayes of 98.00%, with use Support Vector Machine of 76.00% and with c4.5 accuracy obtained pf 95.50%
Keywords: *SMS Classification, Support Vector Machine, Naïves Bayes, C4.5*

1. Pendahuluan

SMS (*Short Message Service*) merupakan fasilitas untuk mengirim atau menerima pesan singka berupa teks melalui telepon genggam. Fasilitas sms ini dapat dinikmati dengan hanya memiliki telepon genggam yang terhubung dengan provider. Fasilitas sms saat ini memang sudah jarang digunakan, tetapi fasilitas ini masih dibutuhkan oleh pengguna telepon genggam. walaupun frekeuensi mengirim ataupun menerima sms saat ini sudah tidak sebanyak dahulu.

SMS terdapat dua macam yaitu sms spam dan sms ham. SMS ham yaitu sms yang berisikan pesan yang benar dan dari seseorang yang dikenal. Sedangkan sms spam yaitu sms yang berasal dari provider atau orang lain yang tidak pernah diminta yang berisi pesan penipuan, penawaran, dan undian.

Saat ini sms spam masih sering diterima oleh pengguna telepon genggam, sms yang berisikan hal-hal yang tidak diinginkan oleh pengguna telepon genggam dapat disebut dengan sms spam. sms berupa spam ini mengganggu para penerima sms bahkan terkadang sms spam ini merugikan orang yang menerimanya dikarenakan sms tersebut berisi penipuan.

Banyaknya sms spam saat ini mengakibatkan para pengguna telepon genggam harus dapat menelaah dan hati-hati terhadap sms yang diterimanya. Tidak sedikit pengguna telepon genggam yang sudah kehilangan uangnya akibat tidak hati-hati saat menerima sms.

Dikarenakan banyaknya sms spam yang diterima, filtering terhadap sms diperlukan untuk membantu para penerima sms mengetahi apakah sms yang diteimanya benar atau tidak. Filtering spam adalah teknik klasifikasi teks yang terbukti menjadi teknik yang hebat untuk mengatasi spam (Sethi & Bhootna, 2014)

Saat ini belum diketahui metode klasifikasi yang akurat dalam mengkalsifikasikan sms, apakah sms yang

Diterima Mei 15, 2017; Revisi Juli 25 , 2017; Disetujui Agustus 16, 2017

diterima berupa spam atau sms tersebut benar. Sehingga perlu diketahui bagaimana akurasi dari metode klasifikasi data mining yaitu Support Vector Machine, Naïves Bayes dan C4.5.

Maksud dari penelitian ini adalah untuk melakukan analisis dan mendapatkan nilai akurat dari komparasi algoritma Support Vector Machine, Naïve Bayes dan C4.5 dalam klasifikasi sms berbahasa Indonesia.

Kajian Literatur

Data Mining "merupakan analisis dari peninjauan kumpulan data untuk menemukan hubungan yang tidak diduga dan meringkas data dengan cara yang berbeda dengan sebelumnya, yang dapat dipahami dan bermanfaat bagi pemilik data." (Larose, 2005)

Klasifikasi merupakan data baru diklasifikasikan didasarkan pada data training. Klasifikasi pada data mining untuk memprediksi *label class* dan mengklasifikasi data didasarkan pada data training dan nilai label class dalam mengklasifikasikan *attribute* dan menggunakan saat mengklasifikasikan data baru.

Langkah dari klasifikasi proses (Han & Kamber, 2006)

1) *Data Cleaning*
2) *Relevance Analysis*
3) *Data transformation and reduction*.

Data mining klasifikasi memiliki beberapa algoritma, yaitu:

a) Decision Tree Classification
b) Naive Bayes Classification
c) Rule-Based Classification
d) Neural Network
e) Support Vector Machines
f) Associative classification
g) K-Nearest-Neighbor Classifiers
h) Genetic Algorithm
i) Rough set Approach
j) Fuzzy Set Approaches

Text Mining "sebagian besar dari informasi yang tersedia disimpan dalam *database* teks (atau *database* dokumen), yang terdiri dari dokumen-dokumen yang besar dari berbagai sumber, seperti artikel berita, makalah penelitian, buku, perpustakaan digital, e-mail, halaman web. *Database* teks yang berkembang pesat karena meningkatnya jumlah informasi yang tersedia secara elektronik bentuk, seperti publikasi elektronik, berbagai macam dokumen elektronik , e-mail , dan WWW (yang juga dapat dilihat sebagai besar, saling berhubungan, dinamis teks *database*)." (Han & Kamber, 2006)

Klasifikasi SMS "Masalah dari klasifikasi dokumen atau pesan (sms) adalah dari konten mereka, sebagai contoh,pesan berupa spam atau bukan spam. Dokumen adalah penggambaran dari set dokumen (spam atau bukan spam) yang mana dapat membuat model seperti kata-kata". (Sethi & Bhootna, 2014)

a. Naïves Bayes

"Bayesian klasifikasi adalah pengklasifikasi statistik yang dapat digunakan untuk memprediksi probabilitas keanggotaan suatu class. Bayesian classification didasarkan pada teorema bayes yang memiliki kemampuan klasifikasi serupa dengan Decision Tree dan Neural Network.Bayesian classification terbukti memiliki akurasi dan kecepatan yang tinggi saat diaplikasikan kedalam *database* dengan data yang besar." (Kusrini, 2009).

Teorema Bayes memiliki bentuk umum sebagai berikut:

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)}$$

Yang mana:

X =data dengan class yang belum diketahui
H = hipotesis data X merupakan suatu class spesifik
P(H|X) = probabilitas hipotesis H berdasar kondisi X (posteriori probability)
P(H) = probabilitas hipotesis H (prior probability)
P(X|H) = probabilitas X berdasar kondisi pada hipotesis H
P(X) = probabilitas dari X

b. Support Vector Machine

"Support Vector Machine merupakan *machine learning* yang termasuk dalam model *supervised* learning atau pembelajaran dengan pengawasan yang berhubungan dengan analisis data dan pengenalan pola." (Susanto & Setiyawan, 2015)

Fungsi keputusan klasifikasi sign (f(x)):

$$f(x) = wx + b$$

$$atau$$

$$fx = \sum_{i=1}^{m} a_i y_i K(x, x_i) + b$$

Keterangan:
N : Banyaknya data
n : dimensi data atau banyaknya fitur
Ld : Dualitas Lagrange Multipier
αi : nilai bobot setiap titik data
C: nilai konstanta
M :jumlah support vector/titik data yang memiliki αi > 0
K(x,xi) : fungsi kernel

c. C4.5

"Merupakan metode klasifikasi yang melibatkan konstruksi pohon keputusan, koleksi node keputusan, terhubung oleh cabang-cabang, memperpanjang bawah dari simpul akar sampai berakhir di node daun." (Sukardi &Supriyanto,2014) .

Tahapan dalam membuat sebuah pohon keputusan dengan algoritma C4.5.

1) Mempersiapkan data training
2) Menghitung total entropy sebelum di cari masing-masing entropy class

$$H(T) = -\sum_{j} Pj\ log2(Pj)$$

Keterangan:
H: Himpunan kasus
T: Atribut
Pj: proposi dai Hj terhadap H

3) Hitung nilai Gain dengan information gain dengan rata-rata

$$Gain\ average = H(T) - Hsaving(T)$$

Keterangan:
H(T)=Total Entropy
Hsaving(T)=Total Gain information untuk masing-masing atribut

Tinjauan Studi

a. Model Penelitian Ika Novita Dewi dan Catur Supriyanto

Penelitian yang dilakukan oleh Dewi dan Catur mengenai klasifikasi SMS spam menggunakan algoritma Naïves Bayes, didalam penelitian ini mendapatkan akurasi 84,40%. Data yang digunakan dalam pengujian yaitu dengan menggunakan data dari UCI Machine Learning Repository dengan data berjumlah 5.574. langkah-langkah yang dilakukan dalam penelitian ini yaitu proses dokumen berupa *tokenize*, *filter stopword* dan *stem*.langkah berikutnya yaitu berupa *wordcreation* dan *purner*.

b. Model Penelitian Tej Bahadur Shahi dan Abhimanu Yadav

Penelitian yang dilakukan oleh Tej Bahadur Shahi dan Abhimanu Yadav mengenai membandingkan 2 metode yaitu Naïves bayes dan Support Vector Machine dalam mengklasifikasikan SMS. Dalam penelitian ini dilakukan 7 kali training data, training data yang pertama dengan 10 sms mendhasilkan tingkat akurasi 80% untuk support vector machine dan 90% untuk Naïves Bayes.

Langkah-langkah yang dilakukan terdiri dari proses prepocessing untuk sms , yang berguna untuk menyeragamkan format agar dapat dimengerti. Langkah selanjutnya yaitu TF-IDF *Calculation* and *Feature Vector Construction*. Dan langkah yang terakhir yaitu klasifikasi.

c. Model Penelitian Sukardi, Abd Syukur dan catur Supriyanto

Penelitian ini mengenai klasifikasi sms email menggunakan algoritma c4.5 dengan seleksi fitur., didalam penelitian ini nilai akurasinya 92.46% yang menggunakan seleksi fitur Information Gain. Data yang digunakan mengambil data dari UCI Repository of Machine Learning.dengan total email 4601 email yang terdiri 2788 berupa non spam dan 1813 berupa email spam. dalam pengolahan data ini dengan menggunakan information gain dengan ratio nilai p=0.6.

Tabel 1. Perbandingan Penelitian Terkait

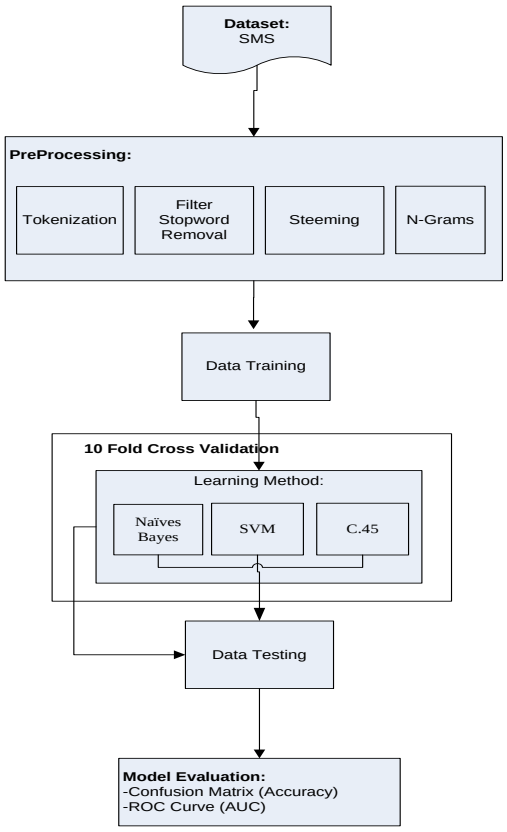
Judul	Classifier	Seleksi Fitur	Accuracy
Mobile sms spam filtering for Nepali text using Naives Bayesian and Support Vector Machine (Shahi dan Yadav, 2014)	Naives Bayes	-	95.33 %

Klasifikasi teks pesan spam menggunakan algoritma Naives Bayes (Dewi dan Supriyanto, 2013)	Naïve Bayes	-	84.40
Mobile sms spam filtering for Nepali text using Naives Bayesian and Support Vector Machine (Shahi dan Yadav, 2014)	Support Vector Machine	-	92.67 %
Klasifikasi Spam Email Menggunakan Algoritma C4.5 dengan seleksi fitur (Sukardi, Syukur dan Supriyanto, 2014)	C.45	Information Gain	92.46 %
Komparasi Algoritma Support Vector Machine, Naïve Bayes dan C4.5 untuk Klasifikasi SMS	Naïve Bayes, SVM dan C4.5	-	?

2. Metode Penelitian

Metode penelitian eksperimen, dengan tahapan berikut:

1. Pengumpulan Data
   Pengumpulan data menggunakan 100 data sms spam dan 100 data sms non-spam.
2. Pengolahan Awal Data
   Dataset ini dalam tahap *preprocessing* harus melalui 4 proses, yaitu:
   a) Tokenisasi
   Dalam proses ini, kata yang memiliki tanda baca dihilangkan, serta dihilangkan juga apabila terdapat simbolyang bukan huruf.
   b) Stopword Removal
   Proses penghapusan atau pembuangan kata-kata yang sering ditampilkan dalam dokumen.
   c) Steeming
   Proses pembuangan prefix dan suffix sehingga membentuk menjadi kata dasar.
   d) N-grams
   Potongan n karakter dalam suatu string tertentu atau potongan n kata dalam suatu kalimat tertentu.
3. Metode yang diusulkan
   Untuk mengetahui metode klasifikasi data mining yang paling akurat pada klasifikasi sms berbahasa Indonesia. Metode yang digunakan yaitu Naïves Bayes, Support Vector Machine dan C4.5.



Gambar 1. Model yang diusulkan

4. Eksperimen dan Pengujian Metode
   Proses eksperimen menggunakan RapidMiner 5.3 untuk pengujian model dilakukan menggunakan dataset.
5. Evaluasi dan Validasi Data
   Validasi dilakukan menggunakan 10 *fold* cross validation. dimana data akan dibagi menjadi 10 bagian.

Eksperimen dan Pengujian Model

Proses eksperimen yang penulis lakukan ini menggunakan RapidMiner 5.3. untuk pengujian model dilakukan menggunakan dataset sms. Tahapan pengujian untuk mengklasifikasi sms sbeagai berikut:

1. Menyiapkan dataset untuk eksperimen yang sudah diketahui classnya
2. Medesain arsitekur algoritma klasifikasi Naïve Bayes, Support Vector Machine dan C4.5
3. Melakukan training dan testing terhadap algoritma Naives Bayes, Support Vector Machine dan C4.5 dan mencatat hasil *accuracy* dan AUC

Evauasi dan Validasi Hasil

Validasi dilakuan menggunakan 10 *fold cross validation*. Untuk 10 *fold cross validation* data eksprimen akan dibagi menjadi 10 bagian. Satu bagian untuk data testing sedangkan Sembilan bagian lainnya untuk data training. Sedangkan pengukuran akurasi diukur dengan *confusion matrix* dan kurva ROC (*Receiver Operating Characteristics*) untuk mengukur nilai AUC.

Tabel 2. *Confusion Matrix*

Classification	**Predicted Class**		
Observed Class		Class = Yes	Class = No
	Class = Yes	a (True Positive – TP)	b (False negative – FN)
	Class = No	c (False Positive – FP)	d (True Negative – TN)

3. Hasil dan Pembahasan

Pada bagian ini, dijelaskan hasil Tujuan dari penelitian ini melakukan analisis dan komparasi untuk memperoleh hasil yang paling akurat dari komparasi algoritma Support Vector Machine, Naïve Bayes dan C4.5 untuk klasifikasi sms berbahasa Indonesia.

3.1. Hasil Eksperimen Menggunakan Naïves Bayes

Hasil yang diperoleh dengan menggunakan algoritma Naïves Bayes adalah nilai *accuracy* = 95.00%. Seperti pada Gambar 2 dari sebanyak 100 data sms yang terdiri dari 100 sms spam dan 100 sms non spam, sebanyak 97 data diprediksi sesuai yaitu spam dan sebanyak 3 data diprediksi spam tetapi ternyata non-spam, 93 data diprediksi sesuai yaitu non-spam dan 7 data diprediksi non-spam tetapi ternyata spam.

Tabel 3.
Confusion Matrix Algoritma Naïves Bayes

Accuracy : 95.00% +/- 3.87% (mikro:95.00%)		
	True NonSpam	**True Spam**
Pred. NonSpam	93	3
Pred. Spam	7	97

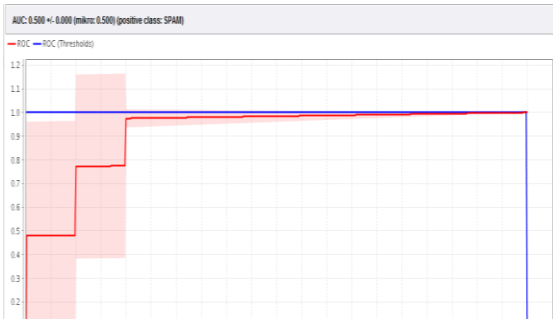
Nilai *accuracy* dari confusion matrix tersebut adalah sebagai berikut:

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

$$accuracy = \frac{(97 + 93)}{(97 + 3 + 93 + 7)}$$

$$accuracy = \frac{190}{200} = 0.95 = 95.00\%$$

Data uji akan dinilai hasil prediksi dengan menggunakan grafik untuk algoritma Naïves Bayes, visualisasi dari grafik ROC dapat dilihat pada Gambar 2.



Gambar 2. Grafik ROC dengan model algoritma Naïves Bayes

Pada Gambar 2, dapat dilihat kinerja algoritma Naïves Bayes mendekati titik 0,1 sehingga dapat dilihat kinerja algoritma ini bagus.

3.2. Hasil Eksperimen Menggunakan Support Vector Machine

Hasil yang diperoleh dengan menggunakan algoritma Support Vector Machine adalah nilai *accuracy* = 76.00%. Seperti pada Gambar 4 dari sebanyak 100 data sms yang teridi dari 100 sms spam dan 100 sms non spam, sebanyak 52 data diprediksi sesuai yaitu spam dan sebanyak 48 data diprediksi spam tetapi ternyata non-spam, 100 data diprediksi sesuai yaitu non-spam dan 0 data diprediksi non-spam tetapi ternyata spam.

Tabel 4. Confusion Matrix Algoritma Support Vector Machine

Accuracy : 76.00% +/-8.31% (mikro: 76.00%)		
	True NonSpam	**True Spam**
Pred. NonSpam	100	48
Pred. Spam	0	52

Nilai accuracy dari confusion matrix tersebut adalah sebagai berikut:

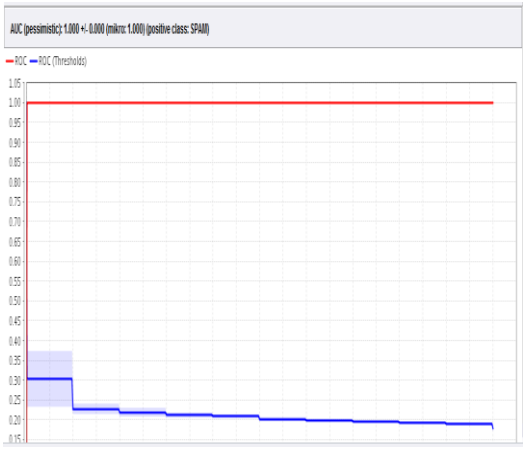
$$accuracy = \frac{(TN + TP)}{(TN + FN + TP + FP)}$$

$$accuracy = \frac{(52 + 100)}{(52 + 48 + 100 + 0)}$$

$$accuracy = \frac{152}{200} = 0.76 = 76.00\%$$

Data uji akan dinilai hasil prediksi dengan menggunakan grafik untuk algoritma Support Vector Machine, visualisasi dari grafik ROC dapat dilihat pada Gambar 3.

Dari Gambar 3 dapat dilihat kinerja algoritma Support Vector Machine agak menjauh dari titik 0,1 sehingga dapat diketahui kinerja algoritma ini tidak lebih baik dari algoritma Naïves Bayes.



Gambar 3. Grafik ROC dengan model algoritma Support Vector Machine

3.3. Hasil Eksperimen Menggunakan C4.5

Hasil yang diperoleh dengan menggunakan algoritma Support Vector Machine adalah nilai *accuracy* = 95.50%. Seperti pada Gambar 6 dari sebanyak 100 data sms yang teridi dari 100 sms spam dan 100 sms non spam, sebanyak 92 data diprediksi sesuai yaitu spam dan sebanyak 8 data diprediksi spam tetapi ternyata non-spam, 99 data diprediksi sesuai yaitu non-spam dan 1 data diprediksi non-spam tetapi ternyata spam.

Tabel 5. Confusion Matrix Algoritma C4.5

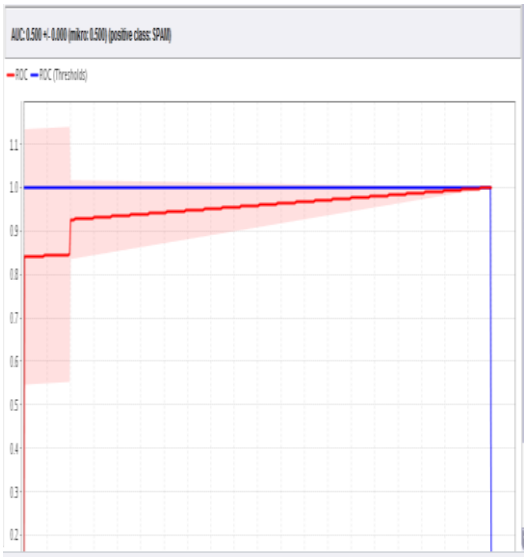
Accuracy : 95.50% +/- 5.68% (mikro:95.50)		
	True NonSpam	**True Spam**
Pred. NonSpam	99	8
Pred. Spam	1	92

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

$$accuracy = \frac{(92 + 99)}{(92 + 8 + 99 + 1)}$$

$$accuracy = \frac{190}{200} = 0.955 = 95.50\%$$

Data uji akan dinilai hasil prediksi dengan menggunakan grafik untuk algoritma C4.5, visualisasi dari grafik ROC dapat dilihat pada Gambar 4.



Gambar 4.
Grafik ROC dengan model algoritma C4.5

Dari Gambar 4 dapat dilihat kinerja algoritma C4.5 lebih mendekati dari titik 0,1 sehingga dapat diketahui kinerja algoritma ini lebih baik dari algoritma Naïves Bayes dan Support Vector Machine.

4. Kesimpulan

Kesimpulan yang diperoleh setelah melakukan analisis dan komparasi untuk memperoleh hasil yang paling akurat pada klasifikasi sms berbahasa Indonesia dengan menggunakan algortima *Support Vector Machine*, *Naives Bayes* dan C4.5.

Dapat dilihat dari hasil pengklasifikasian dengan menggunaan 3 metode aplikasi data mining untuk sms berbahasa Indonesia dengan jumlah data sms 200. Akurasi yang didapat dengan menggunakan Naïves Bayes yaitu sebesar 95.00%, sedangkan yang menggunakan Support Vector Machine sebesar 76.00% dan dengan C4.5 akurasi didapat sebesar 95.50%.

Dapat dilihat dari hasil pengolahan data untuk pengklasifikasian dengan Naïves Bayes, Support Vector Machine dan C4.5. yang paling tinggi akurasinya adalah C4.5 dengan akurasi sebesar 95.50%.

Pada bagian ini, penulis memberikan saran-saran, untuk pengklasifikasian dengan 3 metode klasifikasi : Naïves Bayes, Support Vector Machine dan C4.5, ada beberapa hal yang dapat ditambahkan untuk analisis selanjutnya:

1. Menggunakan metode lain dan ditambah agar hasilnya bias dibandingkan dengan metode yang sudah di uji coba. Baik penggunaan metode-metode terpisah ataupun digabung.
2. Dapat dibuatkan aplikasi untuk *smartphone* agar dapat membantu para penerima sms mengetahui sms yang diterimanya dan dapat meminimalisasi kejadian-kejadian yang tidak diinginkan oleh penerima sms.

Referensi

Dewi, I. N., & Supriyanto, C. (2013). Klasifikasi Teks Pesan Menggunakan Algoritam Naives Bayes. *Seminar Nasional Teknologi Informasi & Komunikasi Terapan 2013 (SEMANTIK 2013).* 156-160.

Han, J., & Kamber, M. (2006). *Data Mining Concepts and Techniques.* San Fransisco: Elsevire.7 & 641.

Kusrini, & E.T, L. (2009). *Algoritma Data Mining.* Yogyakarta: Andi Offset. 189

Larose, D.T. Discovering Knowledge in Data. Canada: Wiley-Interscience, 2005. 2.

Sethi, G., & Bhootna, V. (2014). SMS Spam Filtering Application USing Android . *International Journal of Computer Science and Information Technologies Vol. (5)*.4624-4626.

Shahi, T. B., & Yadav, A. (2013). Mobile SMS SPam Filtering for Nepali text Using Naives Bayesian and Support Vector Machine . *Internationa Journal of Inttelligence*, 24-28.

Sukardi, Syukur, A., & Supriyanto, C. (2014). Klasifikasi Spam Email dengan menggunakan ALgoritma C4.5 dengan Seleksi Fitur. *Jurnal Teknologi Informasi Vol.10 No.1*, 19-30.

Susanto, C. P., & Setiyawan, E. I. (2015). Algoritma Support Vestor Machine Untuk Mendeteksi SMS Spam Berbahasa Indonesia. *Seminar Nasional "Inovasi dalam desain dan Teknologi:*, 109-116.