ALGORITMA NAïVE BAYES DAN METODE UNTUK KLASIFIKASI SMS BERBAHASA INDONESIA

RINGKASAN TESIS

RETNO SARI

14000661

PROGRAM PASCASARJANA MAGISTER ILMU KOMPUTER

SEKOLAH TINGGI MANAJEMEN INFORMATIKA DAN KOMPUTER

NUSA MANDIRI

JAKARTA

2015

HALAMAN PENGESAHAN

Tesis ini diajukan oleh :

Nama	: Retno Sari
NIM	: 14000661
Program Studi	: Magister Ilmu Komputer
Jenjang	: Strata Dua (S2)
Konsentrasi	: *Management Information System*
Judul Tesis	: ”Algoritma Naïve Bayes dan Metode AdaBoost untuk Klasifikasi SMS Berbahasa Indonesia”

Telah berhasil dipertahankan dihadapan Dewan Penguji dan diterima sebagai bagian persyaratan yang diperlukan untuk memperoleh gelar Magister Ilmu Komputer (M.Kom) pada Program Pascasarjana Magister Ilmu Komputer Sekolah Tinggi Manajemen Informatika dan Komputer Nusa Mandiri (STMIK Nusa Mandiri).

Jakarta, 11 Maret 2015
Pascasarjana Magister Ilmu Komputer
STMIK Nusa Mandiri
Direktur

Prof. Dr. Ir Kaman Nainggolan, MS

D E W A N P E N G U J I

Penguji I : Dr. Sularso Budilaksono .....

Penguji II : Dr. Windu Gata, M.Kom .....

Penguji III / Pembimbing : Dana Indra Sensuse, M.LIS, Ph.D .....

STMIK NUSA MANDIRI	**LEMBAR KONSULTASI BIMBINGAN TESIS**
	PASCASARJANA MAGISTER ILMU KOMPUTER STMIK NUSA MANDIRI

Nama : Retno Sari
NIM : 14000661
Dosen Pembimbing : Dana Indra Sensuse, M.LIS, Ph.D
Judul Tesis :Algoritma Naïve Bayes dan Metode AdaBoost untuk Klasifikasi SMS Berbahasa Indonesia

No	Tanggal Bimbingan	Materi Bimbingan	Paraf dosen Pembimbing
1	28 November 2014	Konsultasi Judul	
2	5 Desember 2014	Pengajuan Bab I	
3	11 Januari 2015	Pengajuan Bab II dan Bab III	
4	25 Januari 2015	Pengajuan Bab IV	
5	22 Februari 2015	Pengajuan Bab V	
6	1 Maret 2015	Acc Keseluruhan	

Bimbingan dimulai pada tanggal : 28 November 2014
Bimbingan diakhiri pada tanggal : 1 Maret 2015
Jumlah pertemuan : 6 Pertemuan

Jakarta, 1 Maret 2015

Dosen Pembimbing

(Dana Indra Sensuse, M.LIS, Ph.D)

Halaman

RINGKASAN TESIS

1. PENDAHULUAN

Saat ini sms spam masih belum dapat dihapuskan, setiap orang yang menerima sms harus membaca terlebih dahulu apakah sms yang diterima merupakan spam atau bukan sehingga *filtering* spam sangat dibutuhkan saat ini. *Filtering* spam adalah teknik klasifikasi teks yang terbukti menjadi teknik yang hebat untuk mengatasi spam (Sethi dan Vijender, 2014).

Telah dilakukan penelitian dalam melakukan klasifikasi sms diantaranya, klasifikasi sms dengan Naïves Bayes klasifikasi dan algoritma apriori frequent itemset (Ahmed et al, 2014). *Filtering* sms spam untuk teks berbahasa Nepal menggunakan Naïves Bayes dan Support Vector Machine (Shahi dan Abhimanu, 2014). Teknik *filtering* sms dengan Artifical Immune System (Mahmoud dan Ahmed, 2012). Aplikasi *filtering* sms spam menggunakan android dengan nmenggunakan algoritma Bayesian (Sethi dan Vijender,2014).

Pengklasifikasi Bayes adalah salah satu klasifikasi probabilistik yang sederhana yang mana didasarkan pada teorema Bayes dengan asumsi naïves yang kuat (Ahmed et al, 2014). Selain itu disebutkan pula keuntungan dari klasifikasi Bayes yaitu bahwa Naïves Bayes hanya membutuhkan jumlah yang kecil untuk data training untuk memperkirakan parameter yang dibutuhkan untuk klasifikasi (Korada et al, 2012). Namun, Naïves Bayes memiliki kekurangan yaitu sangat sensitif dalam pemilihan fitur (Chen et al, 2009).

Adaboost digunakan untuk meningkatkan klasifikasi kinerja dari beberapa data yang lemah. Adaboost merupakan algoritma yang paling menjanjikan, konvergensi cepat, dan mudah diimplementasikan kedalam algoritma *machine learning*. Hal ini tidak memerlukan pengetahuan sebelumnya tentang pembelajaran yang lemah dan dapat dengan mudah dikombinasikan dengan metode lain untuk menemukan hipotesis yang lemah, seperti support vector machine (Wang, 2012).

2. LANDASAN/KERANGKA PEMIKIRAN

2.1.Text Mining

Menurut Han dan Kamber (Han dan Kamber, 2006) informasi yang tersedia disimpan dalam *database* teks yang terdiri dari dokumen-dokumen yang besar dari berbaga isumber. Data yang tersimpan di banyak *database* adalah semi struktur data yang mana tidak sepenuhnya tidak tersrtuktur atau terstruktur.

Menurut Bramer (Bramer, 2007) teks merupakan sesuatu yang umum dalam melakukan pertukaran informasi. Syarat umum data dan teks mining adalah informasi yang diambil dan dapat menjadi data yang berguna.Teks mining merupakan proses menganalisa teks untuk menjadi informasi yang berguna untuk tujuan tertentu. Informasi yang diambil harus jelas dan eksplisit, karena teks mining merubah menjad ibentuk yang dapat digunakan oleh *computer* atau orang yang tidak memiliki waktu untuk membaca full teks.

Menurut Charjan, Miss Dipti S dan Pun, Mukesh A. (Charjan, Miss Dipti S dan Pun, Mukesh A , 2013). *Text mining* adalah penemuan dari pengetahuan yang

menarik pada dokumen teks.Hal ini merupakan tantangan untuk menemukan pengetahuan yang akurat pada teks dokumen untuk menolong pengguna untuk menemukan yang diinginkan. Penemuan pengetahuan dapat menjadi efektif digunakan dan memperbaharui pola penemuan dan menerapkannya ke text mining.

2.2.Klasifikasi Data Mining

Klasifikasi pada data mining untuk memprediksi label class dan mengklasifikasi data didasarkan pada daa training dan nilai label class dalam mengklasifikasikan atribut dan menggunakannya saat mengklasifikasikan data baru.

Menurut Han dan Kamber (Han dan Kamber, 2006), langkah dari klasifikasi proses:

a. Data Celaning
b. Relevance analysis
c. Data transformation and reduction

Metode evaluasi klasifikasi menurut Han dan Kamber (Han dan Kamber, 2006);

a. Akurasi, memperkirakan label class
b. Kecepatan, waktu untuk membangun model dan waktu dalam menggunakan model
c. Kehandalan, mengatasi noise dan missing values

2.3.Algoritma Naïve Bayes

Menurut Kusrini dan Luthfi (Kusrini dan Luthfi, 2009) Bayesian klasifikasi adalah pengklasifikasi statistik yang dapat digunakan untuk memprediksi probabilitas keanggotaan suatu class.

Menurut Han dan Kamber (Han dan Kamber, 2006) tahapan dalam algoritma Naïves Bayes:

a. Perhatikan D adalah record training dan ditetapkan label-label kelasnya dan masing-masing record dinyatakan n atribut (n field) X = (X_1, X_2, ..., X_n)
b. Misalkan terdapat m kelas C_1, C_2, ..., C_m
c. Klasifikasi adalah diperoleh maximum posteriori yaitu maximum P(Ci|X)
d. Ini diperoleh dari teorema Bayes

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)} \quad \text{.....(1)}$$

Karena P(X) adalah konstan untuk semua kelas, hanya perlu dimaksumumkan

$$P(C_i|X) = P(X|C_i)P(C_i) \quad \text{.....(2)}$$

2.4.AdaBoost

Menurut Wu dan Kumar (Wu dan Kumar, 2009)AdaBoost merupakan salah satu 10 metode tebaik dari metode data mining. AdaBoost memiliki hasil yang lebih baik dari hasil weak leaner dibandingkan akurasi strong learner yang berubah-ubah.

AdaBoost telah menghasilkan banyak penelitian di aspek teroritical dari *ensemble methods* pada machine learning dan statistik.

Menurut Li, Wang dan Sung (Li, Wang dan Sung, 2008) ditemukannya klasifikasi untuk akurasi yang tinggi dari kombinasi banyak pengklasifikasian komponen yang cukup akurat. Umunya dua teknik yang digunakan untuk membangun ensemble klasifikasi yaitu Boosting dan Bagging.

Metode AdaBoost adalah sebagai berikut:

a. Inisialisasi bobot data {Wn} dengan $Wn^{(m)}$untuk n=1,2,...,N
b. For m=1...,M
   1) Training $Ym^{(x)}$dengan meminimalkan fungsi kesalahan (error function) sebagai berikut:
      $J_m = \sum_{n=1}^{N} W^{(m)_{(n)}} I(Y_m(X_n) \neq t_n$ ..... (3)
   2) Evaluasi kesalahan
      $\varepsilon_m = \frac{\sum_{n=1}^{N} W^{(m)_{(n)}} I(Y_m(X_n) \neq t_n}{\sum_{n=1}^{N} W^{(m)_{(n)}}}$ .....(4)

      Dan kemudian digunakan evaluasi
      $\alpha_m = In\left\{\frac{1-\varepsilon_m}{\varepsilon_m}\right\}$ .....(5)
   3) Memperbaiki (update) bobot data
   $W_n{}^{(m+1)} = W_n{}^{(m)} \exp\{\alpha_m I(Y_m(X_n \neq t_n)\}$ .....(6)
c. Membuat prediksi menggunakan model terakhir sebagai berikut
   $Y_m(X) = sign\left(\sum_{m=1}^{M} \alpha_m Y_{m^{(x)}}\right)$ .....(7)

2.5.*Software Quality Assurance* (SQA)

Kualitas para pembuat *software* dapat dinilai melalui ukuran-ukuran dan metode-metode tertentu, serta melalui pengujian-pengujian *softwar*.

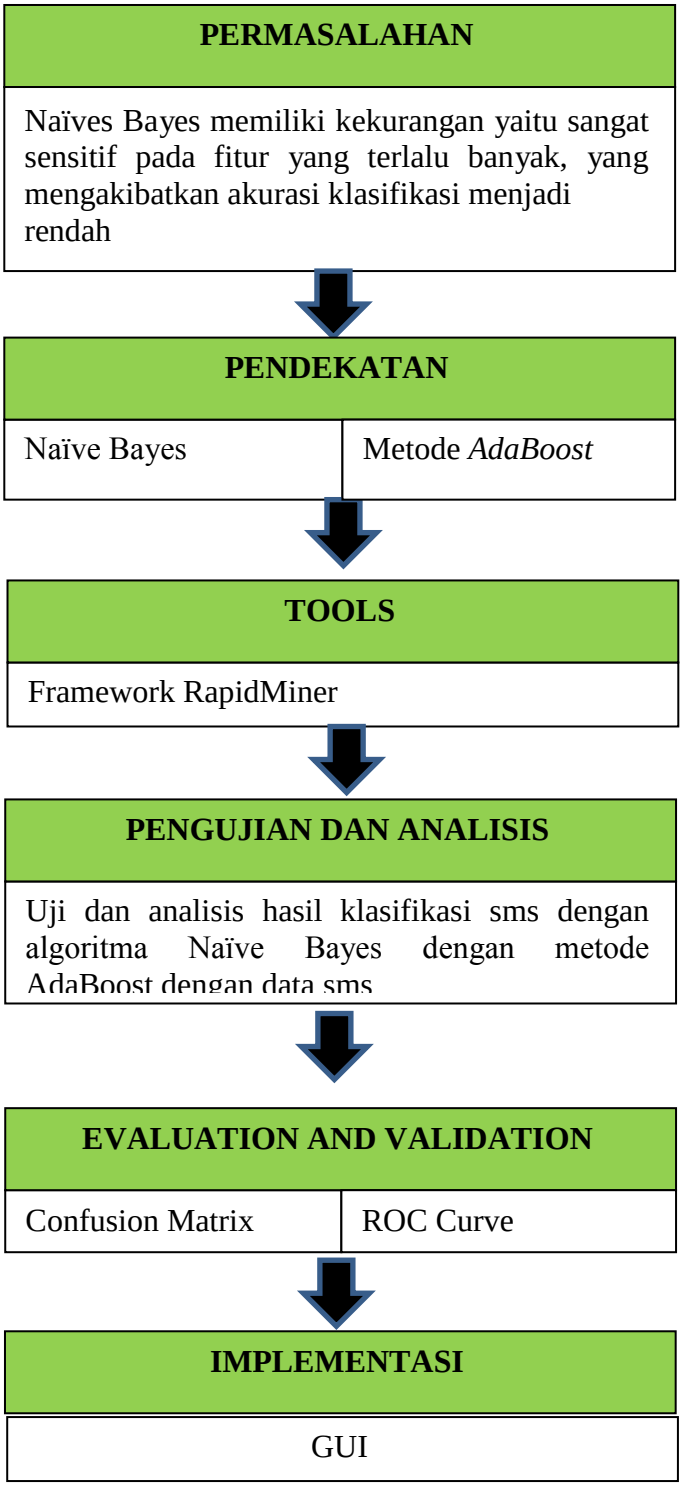
Tabel 1 Metric *Software Quality Assurance* (SQA)

No	Metric	Deskripsi	Bobot
1	*Auditability*	Memenuhi standar atau tidak	0.125
2	*Accuracy*	Keakuratan komputasi	0.125
3	*Completeness*	Kelengkapan	0.125
4	*Error tolerance*	Toleransi terhadap kesalahan	0.125
5	*Excecution efficiency*	Kinerja eksekusi	0.125
6	*Operability*	Kemudahan untuk dioperasikan	0.125
7	*Simplicity*	Kemudahan untuk dipahami	0.125
8	*Training*	Kemudahan pembelajaran fasilitas	0.125

2.6.Kerangka Pemikiran

Penelitian ini adalah mengenai klasifikasi sms dengan menggunakan Naïves Bayes. Dataset yangdigunakan berasal dari sms berbahasa Indonesia yang terdiri dari 250 bukan spam dan 250 spam. Untuk *preprocessing* dilakukan tokenisasi dan Bi-Grams dan teknik boosting yang digunakan adalah AdaBoost, sedangkan

pengklasifikasi yang digunakan adalah Naïves Bayes. Penelitian ini nantinya menghasilkan *accuracy* an pengolahan datanya menggunakan RapidMiner Versi 5.3. Penelitian ini akan dijelaskan secara singkat seperti Gambar 1



Gambar 1 Kerangka Pemikiran

3. METODE PENELITIAN

Metode penelitian yang penulis lakukan adalah metode penelitian eksperimen, dengan tahapan sebagai berikut:

a. Pengumpulan Data

Data dikumpulan berdasarkan data yang akan diproses yaitu berupa sms baik itu sms non-spam maupun sms spam. Data tersebut kemudian diintegrasikan didalam dataset. Pada penelitian ini, hanya mengunakan 250 data sms non-spam dan 250 data sms spam.

b. Pengolahan Awal Data

Untuk mengurangi amanya waktu pengolahan data, penulis hanya menggunakan 250 sms non-spam dan 250 sms spam sebagai data training. Dataset ini dalam tahap *preprocessing* harus melalui 2 proses, yaitu:

1) Tokenisasi

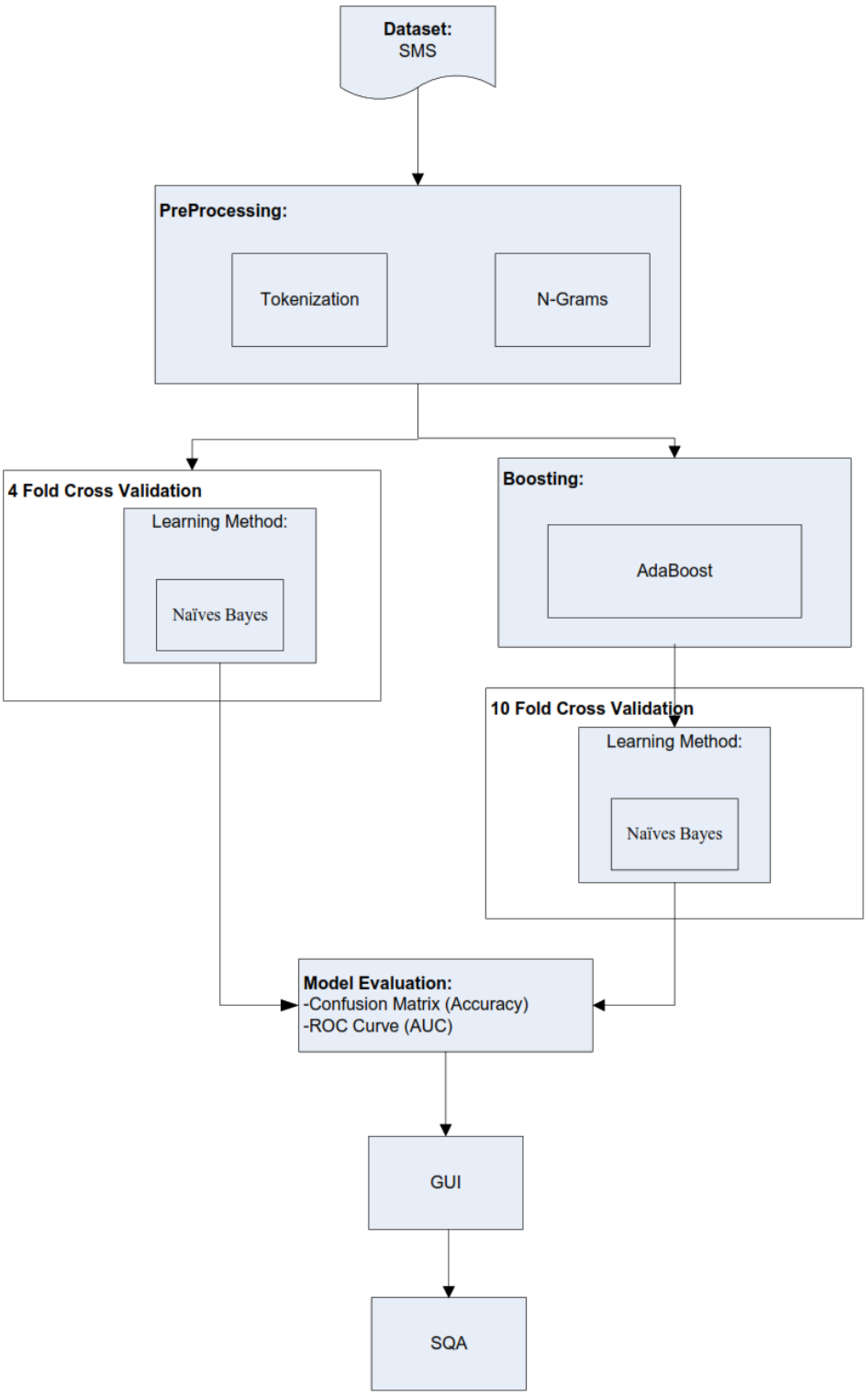
Yaitu proses untuk membagi teks yang dapat berupa kalimat, paragraf atau dokumen menjadi token-token.

2) N-Grams

Yaitu potongan n karakter dalam suatu string tertentu atau potongan n kata dalam suatu kalimat tertentu.

c. Metode Yang Diusulkan

Penambahan metode AdaBoost dilakukan, untuk meningkatkan akurasi pada pengklasifikasi Naïves Bayes. Metode AdaBoost diusulkan untuk meningkatkan akurasi dari pengklasifikasi Naïves Bayes. Pengklasifikasi Naïves Bayes merupakan salah satu algortima yang memiliki akurasi dan kecepatan yang tinggi saat diaplikasikan kedalam *database* dengan data yang besar.



Gambar 2. Metode yang diusulkan

Data harus melalui tahap *preprocessing* terlebih dahulu agar didapatkan kata-kata yang sudah dihilangkan simbol-simbolnya. Proses evaluasi dilakukan menggunakan 4-*fold cross validation* untuk pengujian dengan algoritma Naïves Bayes dan evaluasi dilakukan menggunakan 4-*fold cross validation* untuk pengujian dengan algoritma Naïves Bayes, menggunakan 10-fold *cross validation* untuk pengujian dengan algoritma Naïves Bayes dan AdaBoost. Hasil yang dibandingkan adalah akurasi Naïves Bayes setelah menggunakan metode AdaBoost.

d. Eksperimen dan Pengujian Metode

Eksperimen data dilakukan penulis dengan menggunakan RapidMiner 5 untuk mengolah data. Model diuji untuk melihat hasil yang akan dimanfaatkan dalam mengambil keputusan hasil penelitian.

Tahapan pengujian data untuk mengklasifikasi sms sebagi berikut:

1) Menyiapkan dataset untuk eksperimen yang sudah diketahui classnya
2) Mendesain arsitektur algoritma klasifikasi Naïves Bayes
3) Melakukan training dan testing terhadap algoritma Naïves Bayes dan mencatat hasil accuracy dan AUC
4) Mendesain arsitektur algoritma klasifikasi Naïves Bayes dan teknik boosting AdaBoost
5) Melakukan training dan testing terhadap Naïves Bayes dan AdaBoost dan mencaat hasil accuracy dan AUC

e. Evaluasi dan Validasi Hasil

Validasi dilakukan menggunakan 4 *fold cross validation* dan 10 *fold cross validation*. Untuk 10 *fold cross validation* data eksperimen 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 Characteristic*) untuk mengukur nilai AUC. Dengan *confusion matrix,* akurasi Naïves Bayes sebelum menggunakan metode AdaBoost dan setelah menggunakan metode AdaBoost. Teknik 3.2 berikut adalah tampilan *confusion matrix* dan rumus perhitungannya menurut Gorunescu (Gorunescu,2011):

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 –FN)

$$Akurasi = \frac{a+d}{a+b+c+d} = c = \frac{TP+TN}{TP+FN+FP+TN}$$

4. HASIL PENELITIAN DAN PEMBAHASAN

Dalam proses ini sms diklasifikasikan untuk menentukan class untuk setiap smsnya, class sms terbagi dua yaitu class non-spam dan class spam. Penentuan class untuk setiap sms ditentukan melalui perhitungan probabilitas dari rumus algoritma Naïves Bayes. Class sms diberikan nilai class non-spam apabila nilai probabilitas pada dokumen tersebut untuk nilai class non-spamnya lebih besar dibandingkan dengan class spamnya. Dan suatu sms dikatakan class spam apabila nilai probabilitas pada dokumen tersebut untuk nilai class spamnya lebih besar dibandingkan dengan class non-spamnya. Kehadiran kata di dalam suatu sms akan diwakili oleh angka 1 dan angka 0 jika kata tersebut tidak muncul dari sms.

Tabel 3 Tabel vector dokumen boolean dengan label class hasil klasifikasi

Dokumen	Saya	Ibu	Bapak	Info	Hubungi	Class
NS-1	1	1	0	0	0	Non-spam
NS-2	0	1	1	1	0	Non-spam
NS-3	1	0	0	0	0	Non-spam
S-1	0	1	1	0	1	Spam
S-2	1	0	1	0	1	Spam
S-3	0	0	0	1	1	?

Probabilitas Bayes yang dijabarkan untuk dokumen ke S-3.

a. Hitung probabilitas bersyarat (*likelihood*) dokumen ke S-3 pada class non-spam dan spam untuk class non-spam

P(S-3|Non-spam) = P(Saya=2|Non-spam) x P(Ibu=2|Non-spam) x P(Bapak=1|Non-spam) x P(Hubungi=0|Non-spam) x P(Selamat=1|Non-spam)

P(S-3|Non-spam) = 2/3 x 2/3 x 1/3 x 1/3 x 1/3
= 0.66 x 0.66 x 0.33 x 0.33 x 0
= 0

P(S-3|spam) = P(Saya=1|spam) x P(Ibu=1|spam) x P(Bapak=2|spam) x P(Hubungi=1|spam) x P(Selamat=3|spam)

P(S-3|spam) = 1/2 x 1/2 x 2/2 x 1/2 x 3/2
= 0.5 x 0.5 x 1 x 0.5 x 1.5
= 0.1875

b. Probabilitas prior dari class non-spam dan spam dihitung dengan proporsi dokumen pada tiap class
P(Non-spam)= 3/5= 0.6
P(spam)= 2/5 = 0.4

c. Hitung probabilitas posterior dengan memasukan rumus Bayes dan menghilangkan penyebut (S-3):

$$P(Non-Spam\ |S-3) = \frac{(0)(0.6)}{P(S-3)} = 0$$

$$P(Spam\ |S-3) = \frac{(0.1857)(0.4)}{P(S-3)} = 0.07428$$

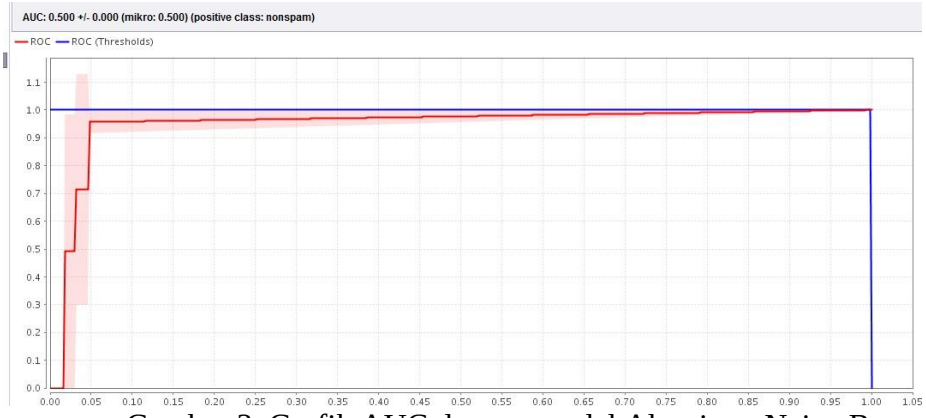
Berdasarkan probabilitas di atas, maka dapat disimpulkan bahwa S-3.txt termasuk dalam class spam, karena P(Non-spam|S-3) lebih kecil dari pada P(spam|S-3).

A. Hasil eksperimen menggunakan algoritma Naïve Bayes

Hasil eksperimen menggunakan algoritma Naïve Bayes di peroleh nilai accuracy = 96.40 % seperti pada tabel 4.4 dan AUC = 0.999. Dari sebanyak 500 data sms yang terdiri dari 250 sms non-spam dan 250 sms spam, sebanyak 243 data diprediksi sesuai yaitu spam dan sebanyak 7 data diprediksi spam tetapi ternyata non-spam, 239 data diprediksi sesuai yaitu non-spam dan 11 data diprediksi non-spam tetapi ternyata spam.

Tabel 4 *Confusion Matrix* Algoritma Naive Bayes

Accuracy :96.40 +/- 2.30% (mikro :96.40%)		
	True NonSpam	**True Spam**
Pred. NonSpam	239	7
Pred. Spam	11	243



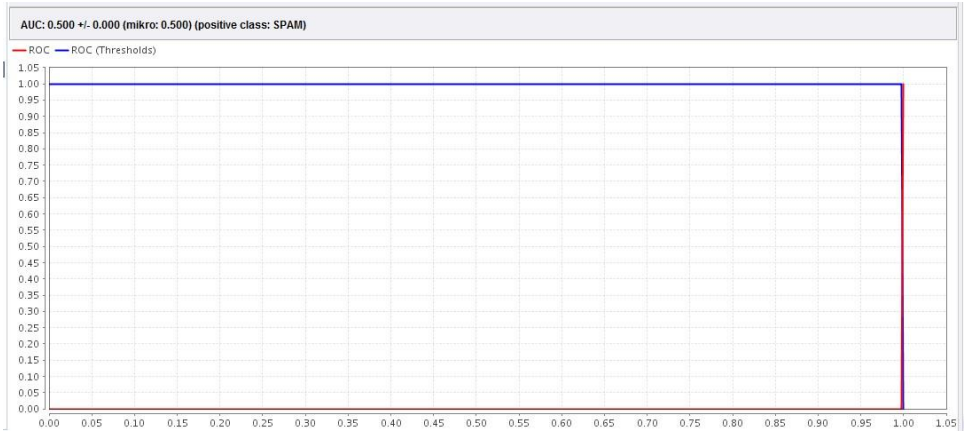
Gambar 3. Grafik AUC dengan model Algoritma Naive Bayes

B. Hasil Eksperimen menggunakan Algoritma Naïve Bayes dan Metode AdaBoost

Hasil eksperimen dengan menggunakan algoritma Naive Bayes setelah ditambahkan metode AdaBoost didapakan hasil akurasi dari pengklasifikasian sms dengan algoritma Naives Bayes dari 250 sms non spam dan 250 sms spam nilai akurasinya sebesar 100%. Nilai akurasi ini mengalami peningkatan sebesar 4.2% dari penggunaan algoritma Naives Bayes.

Tabel 5 *Confusion Matrix* Algortima Naïves Bayes Setelah menambahkan metode AdaBoost

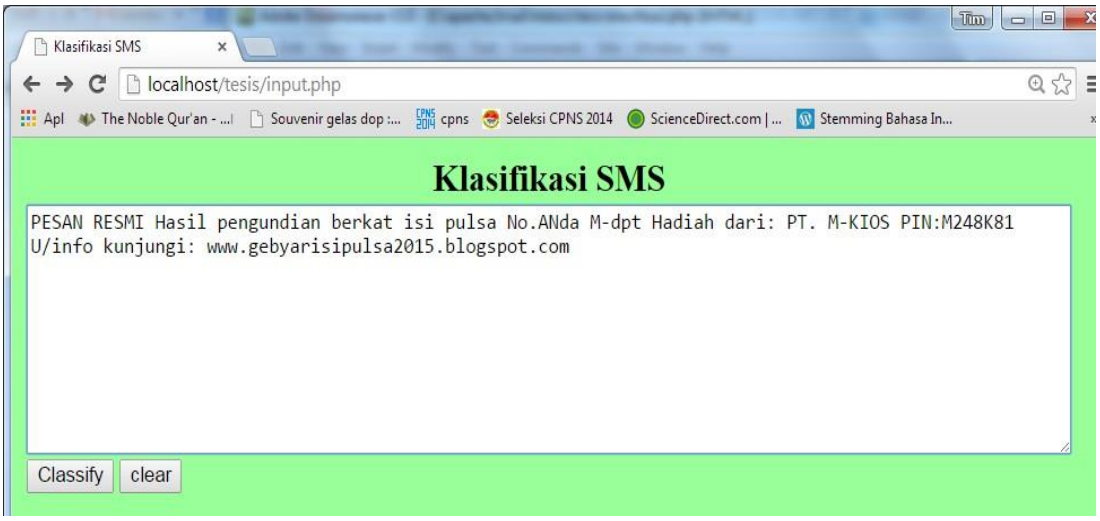
Accuracy : 100 +/- 0.00% (mikro :100.00%)			
	True Negatif	**True Positif**	**Class Precission**
Pred. Negatif	250	0	100.00%
Pred. Positif	0	250	100.00%
Class Recall	100.00%	100.00%	



Gambar 4. Grafik ROC dengan model algoritma Naïves Bayes dan metode AdaBoost

C. Pengembangan Prototipe

Hasil klasifikasi dari penelitian akan diterapkan kedalam pembuatan aplikasi untuk klasifikasi sms dengan menggunakan bahasa pemrograman PHP, sehingga dapat memudahkan para penerima sms untuk mengklasifikasikan sms yang diterima, seperti terlihat dalam gambar dibawah ini:



Gambar 5 Tampilan Aplikasi Mengklasifikasi SMS Spam

Gambar 5 merupakan tampilan aplikasi untuk mengklasifikasi sms, dimana dalam aplikasi tersebut terdapat sms yang diindikasikan sebagai spam. Pada aplikasi tersebut saat ditekan tombol *classify*, maka hasil dari klasifikasi tampil.

Gambar 6. Hasil Tampilan Aplikasi Mengklasifikasi SMS Spam

D. **Penerapan *Software Quality Assurance* (QSA)**

Penerapan SQA dilakukan dengan mengolah data dari angket yang telah disebarkan, hasil angket yang dilakukan terhadap 4 orang penerima sms yaitu Dosen, instruktur, mahasiswa dan pelajar yang berperan sebagai *user*.

Tabel 6 Hasil Evaluasi SQA

User	Skor *Metric*								Skor
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	
#1	70	65	60	70	65	80	75	60	68.1
#2	75	60	60	75	70	85	65	65	69.4
#3	65	79	55	75	65	85	75	60	69.9
#4	60	65	65	70	60	75	70	62	65.9
Skor Rata-Rata									69.4

Di lihat dari table 6 hasil evaluasi SQA didapatkan skor rata-ratanya yaitu 69.4. Skor ini dilihat dari kriteria skala penilaian memiliki nilai yang optimal untuk sebuah perangkat lunak yang memenuhi standar kualitas berdasarkan uji SQA.

Skor metrik dalam penelitian ini berdasarkan hasil kuesioner adalah sebagai berikut:

Tabel 7 Hasil Evaluasi SQA dalam penilaian skala *Likert*

User	Skor *Metrik*							
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
#1	4	4	3	4	4	4	4	3
#2	4	3	3	4	4	5	4	4
#3	4	4	3	4	4	5	4	3
#4	3	4	4	4	3	4	4	4

Tabel 7 menjelaskan nilai dari skala likert untuk setiap metriksnya pada *user*, pada tabel diatas dapat dilihat banyak skor 4 yang menunjukan bahwa prototype yang dibuat optimal.

Berdasarkan dati table 7, dilakukan perhitungan untuk masing-masing metric terhadap banyaknya yang dinilai oleh *user:*

Tabel 8 Hasil Perhitungan Skala Likert Penelitian tiap Metrik

No	**Metriks**	**Skala Penilaian**					**Jumlah Jawaban**
		SS	**S**	**R**	**TS**	**STS**	
1	*Auditability*	0	3	1	0	0	4
2	*Accuracy*	0	3	1	0	0	4
3	*Completeness*	0	1	3	0	0	4
4	*Error Tolerance*	0	4	0	0	0	4
5	*Execution Efficiency*	0	3	1	0	0	4
6	*Operability*	2	2	0	0	0	4
7	*Simplicity*	0	4	0	0	0	4
8	*Training*	0	2	2	0	0	4

Tabel 8 merupakan hasil perhitungan skala likert untuk tiap metriknya, dari tabel tersebut. Dari table diatas untuk metriks *Auditability* untuk nilai setuju 3 dan nilai ragu-ragu 1, metriks *Accuracy* untuk nilai setuju 3 dan nilai ragu-ragu 1, metriks *Completeness* untuk nilai setuju 1 dan nilai ragu-ragu 3, metriks *Error Tolerance* untuk nilai setuju 4, metriks *Execution Efficiency* untuk nilai setuju 3 dan nilai ragu-ragu 1, metriks *Operability* untuk nilai sangat setuju 2 dan untuk nilai setuju 2, metriks *Simplicity* untuk nilai setuju 4 dan metriks Training untuk nilai setuju 2 dan nilai ragu-ragu 2.

5. KESIMPULAN

Untuk mengklasifikasikan teks dengan data berupa sms, salah satu pengklasifikasi yang dapat digunakan adalah pengklasifikasi Naïve Bayes. Hal ini dikarenakan Naïve Bayes sangat sederhana dan efisien. Selain itu Naïve Bayes juga sangat populer digunakan untuk klasifikasi teks dan memiliki perfoma yang baik pada banyak domain.

Dari pengolahan data yang sudah dilakukan yang telah ditambahkan metode Adaboost, terbukti dapat meningkatkan akurasi pengklasifikasi Naïve Bayes. Data sms dapat diklasifikasi dengan baik kedalam bentuk positif dan negatif. Akurasi Naïve Bayes sebelum ditambahkan metode Adaboost mencapai 96.40%. Sedangkan setelah menggunakan Adaboost,akurasinya meningkat hingga mencapai 100%. Peningkatan akurasi mencapai 3.6% untuk mendukung penelitian, penulis mengembangkan aplikasi untuk mengklasifikasi sms spam dan sms non-spam menggunakan bahasa pemrograman php.

Model yang terbentuk dapat diterapkan pada seluruh sms, sehingga dapat dilihat secara langsung hasilnya dalam bentuk spam dan non-spam. Hal ini dapat membantu seseorang untuk mengetahui sms yang diterima itu adalah spam atau bukan spam.

DAFTAR PUSTAKA

Ahmed, Ishtiaq, Guan, Donghai dan Chung , Tae Choong. (2014) SMS classification based on Naives Bayes classifier and Apriori Algorithm Frequent Itemset. International Journal of Machine Learning and Computing Vol. 4 No.2.

Bramer, Max.(2007). Principles of Data Mining. London: Springer.

Charjan, Miss Dipti S dan Pun, Mukesh A. (2013). Pattern Discovery For Text Mining Using Pattern Taxonomy. International Journal of Engineering Trends and Technology. Volume 4 Issue 10. 4550-4555.

Chen, J., Huang, H., Tian, S., dan Qu, Y. (2009). Feature selection for text classification with Naives Bayes. Expert Systems with Application, 36 (3),5432-5435.

Dewi, Ika Novita dan Supriyanto , Catur. (2013) Klasifikasi Teks Pesan Spam Menggunakan Algoritma Naives Bayes. Seminar Nasional Teknologi Informasi & Komunikasi Terapan 2013 (SEMANTIK 2013). ISBN-979-26-0266-6.

Gorunescu, F. (2011). Data Mining Concept Model Technique. Verlag Berlin Heidelberg:Springer.

Han, Jiawei dan Kamber, Michelin. (2006). Data Mining Concepts and Techniques. San Francisco: Elsevire.

Korada, N. K,Kumar, N. S.P., dan Deekshitulu, Y. V. N.H .(2012) Implementation of Naives Bayesian Classifier and Ada-Boost Algorithm Using Maize Expert System. Interntional Journal of Information Sciences and Techniques, 2.

Kusrini dan Luthfi, E.T. (2009). Algoritma Data Mining. Yogyakarta: Andi Offset.

Li, Xunchun, Wang, Lei dan Sung, Eric. (2008). AdaBoost with SVM-based component classifiers. Engineering Applications of Artifical Intelligence 21.785-795.

Mahmoud, Tarek M dan Mahfouz , Ahmed M. (2012). SMS Spam Filtering Technique Based on Artifical Immune System. International Journal of Computer Science Issues. Vol. 9 Issue 2, No 1.589-597.

Sethi, Gaurav dan Bhootna , Vijender. (2014). SMS Spam Filtering Application Using Android. International Journal of Computer Science and Information Technologies Vol. 5 (3). ISSN:0975-9646.

Shahi, Tej Bahadur dan Yadav , Abhimanu. (2013). Mobile SMS Spam Filtering for Nepali Text Using Naives Bayesian and Support Vector Machine. International Journal of Intelligence Science. 24-28.

Sugiyono. (2012). Metode Penelitian Kuantitatif, Kualitatif dan R&D. Bandung:Alfabeta

Teli, Savita Pundalik dan Biradar, Santoshkumar. (2014). Effective Email Classification for Spam and Non-Spam. International Journal of Advanced Reserach in Computer Science and Software Engineering.

Wang, Lipo dan Fu, Xiuju.(2005). Data mining with Computational Intelligence. Verlag Berlin Heidelberg:Springer.

Wang, Ruihu. (2012). AdaBoost for Feature Selection, Classification and Its Relation with SVM, A Review. 2012 International Conference on Solid State Device and Material Science. 800-807.

Wu, Xindong dan Kumar, Vipin.(2009). The Top Tens Algorithms in Data Mining. New York :Taylor & Francis Group, LLC.