ANALISIS KOMPARASI *MACHINE LEARNING* PADA DATA SPAM SMS

Tri Herdiawan Apandi[1)], Castaka Agus Sugianto[2)]
[1)] Manajemen Informatika, Politeknik Negeri Subang
[2)] Teknik Informatika, Politeknik TEDC Bandung
Email: h.apandi@gmail.com [1)], castaka@poltektedc.ac.id[2)]

Abstrak
Spam SMS adalah pesan yang tidak berguna bagi penerima dan sering kali menjadi penyalahgunaan oleh pihak yang tidak bertanggung jawab. Menhindari penyalahgunaan itu perlu dilakukan penyaringan spam SMS, tetapi perlu algoritma penyaringan data spam SMS. Dengan menggunakan *Machine Learning* penyaringan itu menjadi mudah, contoh dari *Machine Learning* yang popular adalah SVM dan *Naïve Bayes*. SVM dan Naïve Bayes dapat digunakan untuk penyaringan data spam SMS, tetapi machine learning mana yang menjadi terakurat dan memiliki nilai presisi yang baik. Untuk melihat komparasi antar kedua algoritma tersebut maka dilakukan cara pengolahan data spam SMS dengan cara mengumpulkan data SMS tertebih dahulu kemudian data SMS tersebut diberi label manual lalu dilakukan proses akromin, *stop words* dan pembobotan. Setelah dilakukan pembobotan maka akan dilakukan proses training oleh SVM dan *Naïve Bayes*. Proses training dilakukan untuk mendapatkan model yang akan diuji untuk menbandingkan machine leaning pada data Spam SMS. Setelah dilakukan pengujian dengan membuat 12 model data, maka didapat SVM memiliki nilai presisi yang lebih baik dari pada *Naïve Bayes* yaitu 94.98%. *Naïve Bayes* memiliki tingkat akurasi yang baik dengan rata-rata 92.22%.

Kata Kunci: SVM, naïve bayes, spam SMS, n-grams

Abstract
SMS spam is a message that is not useful to the recipient and often becomes abuse by irresponsible parties. Avoiding abuse is necessary spam filtering SMS, but need SMS spam data filtering algorithm. Using Machine Learning filtering becomes easy, examples from the popular Machine Learning are SVM and Naïve Bayes. SVM and Naïve Bayes can be used for SMS spam data filtering, but which machine learning becomes accurate and has good precision values. To see the comparison between the two algorithms, then do the data processing SMS spam by collecting the first SMS data then the SMS data is labeled manually and then performed the process acronyms, stopwords and weighting. After the weighting will be done training process by SVM and Naïve Bayes. The training process is done to get the model to be tested to compare the leaning machine on SMS Spam data. After testing by making 12 data models, then obtained SVM has a precision value better than the Naïve Bayes is 94.98%. Naïve Bayes has a good degree of accuracy with an average of 92.22%

Keywords: *SVM, naïve bayes, spam SMS, n-grams*

I. PENDAHULUAN

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

Fasilitas SMS ini sering disalahgunakan oleh oknum yang tidak bertanggung jawab sehingga efeknya bisa merugikan penerima, ini sering disebut dengan istilah Spam SMS(Anik, 2013). Di tahun 2011 kurang dari 1% dari sms yang terkirim di Amerika Utara merupakan Spam SMS, sedangkan di wilayah Asia sendiri jumlahnya mencapai 30% dari sms yang terkirim mengandung Spam SMS (Hastie & Tibshirani, 2009). Hal ini di karenakan rendahnya filtering SMS yang masuk, sehingga meningkatnya jumlah *spam* SMS dan banyak *spammers* yang muncul. Untuk itu proses penyaringan SMS sangat dibutuhkan, terlebih untuk proses pencegahan atau meminimalisir kerugian yang dihadapi oleh masyarakat umum pengguna telepon seluler khususnya pengguna layanan pesan singkat ini, dan juga *provider*. Pada penelitian sebelumnya mengenai penyaringan spam SMS banyak metode yang digunakan salah satunya yang menggabungkan *Support Vector Machine* dan token memiliki kekurangan pada saat isi dari SMS menggunakan imbuhan hasilnya tidak maksimal (Hastie & Tibshirani, 2009). *Naïve Bayes* (Anon, 2012), algoritma ini memberikan hasil bahwa waktu pemrosesannya dan belajar lebih cepat serta tingkat akurasi yang wajar dibandingkan

dengan algoritma *Decision Trees* dan *sampling algorithm* (Hu & Yan, 2010). *Neural Network* memberikan *error* generalisasi yang lebih besar dibandingkan dengan *Support Vector Machine* (Almeida et al., 2011). Banyaknya jumlah fitur dan proses pemecah kalimat menjadi kata mempengaruhi tingkat akurasi yang dihasilkan, pada penelitian dilakukan pemecahan kata berdasarkan pemilihan 1 kata saja. Sehingga kata "selamat" yang terdapat di spam SMS dan kata "selamat" yang terdapat ham SMS mempunyai bobot nilai yang sama, sehingga akan membingungkan *machine learning* (Apandi & Sugianto, 2015). Untuk membedakan kata yang sama dengan makna yang berbeda, diperlukan pencacah kata yang salah satunya menggunakan N-Grams. N-Gram dapat berfungsi untuk pengambilan potongan n karakter dalam suatu string atau kalimat tertentu(Preoţiuc Pietro & Florentina, 2012).

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. untuk menentukan posisi optimal dari *hyperplane* di *dual space* pada algoritma *Support Vector Machine (SVM)* dilakukan ketika proses *training*, sedangkan untuk memisahkan jarak antara fungsi pemisah (*separating hyperplane*) ke masing-masing kelas menggunakan margin. SVM pada saat proses *training* mencari *training* set yang paling sempurna. Sedangkan pada metoda yang lain, pada proses training untuk menemukan *local* optimal perlu dilakukan secara berulang (Srivastava & Bhambhu, 2010). Untuk mendapatkan nilai yang optimal algoritma SVM pada saat *training* waktu yang diperlukan hanya sekali, sehingga halini mengurangi terjadinya *overfitting* karena disebabkan *overtrained.* Berdasarkan hasil penelitian terdahulu, dapat ditarik kesimpulan bahwa algoritma SVM memiliki kinerja dasar terbaik (Almeida et al., 2011). Penelitian ini bertujuan untuk mengetahui hasil perbandingan akurasi dari tokenisasi kata N-Gram antara menggunakan algoritma *Support Vector Machine (SVM)* dan *Naïve bayes*.

II. LANDASAN TEORI

Filtering

Proses pemilahan secara otomatis mana email atau sms yang "benar" dan mana spam dengan ini bisa menghemat waktu dan tenaga disebut filtering. Permasalahan spam yang marak ini memicu banyaknya solusi pemfilteran dari yang paling sederhana sampai yang paling kompleks (Saini & Desai, 2012). Dari pemfilteran bersifat personal hingga masal dan dari yang geratisan hingga berbayar.

Short Message Service

Kemampuan untuk mengirim dan menerima pesan dalam bentuk teks dari dan kepada ponsel disebut *Short Message Service* (SMS). Teks yang di dalam SMS tersebut bisa terdiri dari kata-kata atau nomor atau kombinasi keduanya. *Europesan Telecommunication Standards Institute* (ETSI), menciptakan standar pesan (*message*) SMS, ETSI juga membuat standar GSM yang sampai sekarang diimplementasikan oleh semua operator GSM. Pada bulan Desember 1992 di UK SMS pertama yang dikirimkan dari PC ke sebuah ponsel melalui jaringan GSM Vodafone. Pada aturannya jika menggunakan alphabet Latin Setiap Pesan maksimal terdiri dari 160 karakter dan jika menggunakan alphabet non-Latin seperti huruf Arab atau China terdiri dari 70 karakter (Brown et al., 2007). *Short Message Service* pada sistem komunikasi tanpa kabel (*wireless*) merupakan sebuah layanan yang banyak diaplikasikan, pengiriman pesan memungkinkan dilakukan dalam bentuk alphanumeric dengan sistem eksternal antara terminal pelanggan atau antar teminal pelanggan, seperti *e-mail*, paging, *voice* mail, dan lain-lain.

Spam SMS

Mengirimkan pesan secara bertubi-tubi tanpa dikehendaki oleh penerimanya menggunakan perangkat elektronik yaitu adalah spam (Jane et al., 2012). Tindakan ini dikenal dengan nama spamming. Banyak spam yang dikenal secara umum bentuk spam meliputi : spam surat elektronik, spam Usenet newsgroup, spam pesan instan, spam wiki, spam mesin pencari informasi web (web search engine spam), spam blog, spam jejaring social dan spam iklan baris daring (Jane et al., 2012).

Support Vector Machine

Support vector machine (SVM) adalah algoritma yang sangat populer belakangan ini merupakan suatu teknik yang relatif baru (1995) baik dalam kasus klasifikasi maupun regresi untuk melakukan prediksi (Hastie & Tibshirani, 2009). SVM dan ANN keduanya masuk dalam kelas *supervised learning*, baik dalam hal fungsi dan kondisi permasalahan yang bisa diselesaikan. SVM memberi hasil yang lebih baik dari ANN, terutama dalam hal solusi yang dicapai, ini terbukti dalam banyak implementasi, SVM menemukan solusi yang global optimal sedangkan ANN menemukan solusi berupa *local optimal*. Ketika menjalankan ANN solusi dari setiap training suka berbeda, ini tidak heran karena hal ini penyebabnya solusi *local optimal* yang dicapai oleh ANN tidak selalu sama. Berbeda dengan SVM yang selalu mencapai solusi yang sama setiap *running*. Dalam teknik ini, berusaha memisahkan dua set data dari dua kelas yang berbeda dengan cara menemukan fungsi pemisah (*classifier*) yang paling optimal (Hastie & Tibshirani, 2009)

A. Naïve Bayes

Naïve bayes classifier (NBC) adalah salah satu algoritma yang ada dalam teknik data mining yang menerapkan teori Bayes dalam proses klasifikasi (Santoso, 2007). *Naïve bayes classifier* merupakan pengklasifikasian statistik yang dapat digunakan untuk memprediksi probabilitas keanggotaan suatu class. *Naïve Bayes* sudah terbukti memiliki nilai akurasi yang tinggi dan kecepatan yang cepat jika diimplementasikan ke dalam *dataset* yang besar.

Teorema Bayes memiliki bentuk umum sebagai berikut.

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

Keterangan:

- 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

Dalam terminologi sederhana, algoritma naïve bayes mengasumsikan bahwa kehadiran (atau ketiadaan) fitur tertentu dari suatu kelas tidak berhubungan dengan kehadiran (atau ketiadaan) fitur lainnya. Sebagai contoh, buah mungkin dianggap apel jika merah, bulat, dan berdiameter sekitar 4 inchi. Bahkan jika fitur ini memiliki ketergantungan satu sama lain atau atas keberadaan fitur lain, seluruh sifat-sifat berkontribusi mandiri untuk probabilitas bahwa buah ini adalah apel menurut algoritma *naïve bayes*. Tergantung pada situasi yang tepat dari model probabilitas, dalam *supervised learning algoritma* NB dapat dilatih sangat efisien. *Naive bayes* menyederhanakan perhitungan probabilitas dengan mengasumsikan bahwa probabilitas setiap atribut yang termasuk dalam nilai kelas tertentu tidak memiliki ketergantungan pada semua atribut lainnya. Probabilitas bersyarat adalah probabilitas nilai kelas yang diberi nilai atribut. Dengan mengalikan probabilitas kondisional bersama untuk setiap atribut untuk nilai kelas tertentu, kita memiliki probabilitas *instance* data yang termasuk dalam kelas tersebut (Brownlee, 2014).

III. METODE PENELITIAN

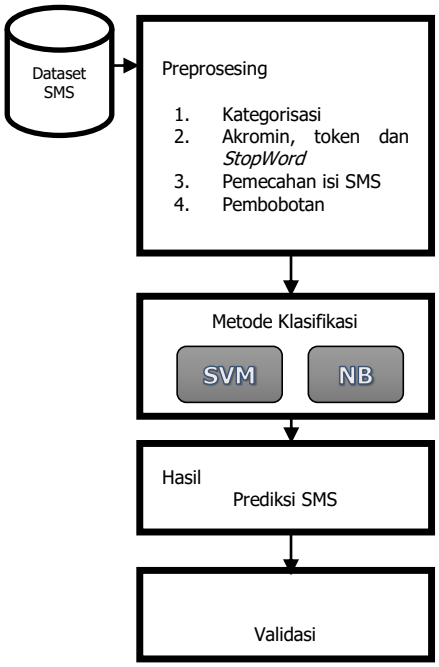
Model yang disajikan bias dilihat pada **gambar 1** prosesnya dimulai dari pengumpulan *dataset* sampai validasi. Pada fase pertama ini *dataset* dilakukan proses kategorisasi, akromin, tokenisasi, *stop word*, pemecahan isi sms dan pembobotan, pada tahap berikutnya hasil dari pengkategorian akan dikelasifikasikan dengan metode *Support Vector Machine* dan *Naïve Bayes*, dan fase terakhir akan divalidasi.

a. *Dataset*: *dataset* yang dikumpulkan sebanyak 900 *record*, baik berupa spam sms dan yang bukan spam.
b. Kategorisasi: proses ini menentukan mana sms yang *spam* dan yang bukan/ham. Dengan memberi label *spam* dan *non spam*. Cara untuk membedakan spam dan bukan adalah dengan cara memberikan label dari tiap - tiap data. Data yang spam diberi label *spam* dan yang bukan/ham diberi label *not spam*.
c. Akromin: merupakan proses mengganti singkatan dengan kata bakunya. Contoh dari tahap ini adalah "jt" menjadi "juta", "bb" menjadi "blackberry", dan "dg" menjadi "dengan".
d. *Stopwords*: pada proses ini dilakukan proses membuang kata–kata sambung yang sering muncul dan tidak bermakna apa-apa. Maksud tidak bermakna apa-apa contohnya dapat dilihat di tabel 1.

Tabel 1. Proses *stop words*

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.

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 *Binary Term*, TF-IDF dan *Term Frequency* (TF).
h. Metoda klasifikasi: metode yang digunakan adalah *Support Vector Machine* (SVM) dan *Naïve Bayes* (NB) untuk klasifikasi. Pada klasifikasi menggunkan *Support Vector Machine* dan *Naïve Bayes* ini akan menghasilkan model yang dipakai untuk menguji data testing.
i. Hasil prediksi: hasil prediksi dari data training berupa akurasi dari tiap-tiap model.
j. Validasi: untuk memvalidasi hasil pengujian menggunakan *Cross-validasi*.



Gambar 1. Model yang diusulkan

IV. 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 *Binary Term*, TF-IDF dan *Term Frequency* (TF). Kemudian dilakukan kelasivikasi, setelah itu akan dilihat model mana yang memiliki akurasi yang tinggi dari hasil proses pengujian yang dilakukan.

Pengujian

Dalam pengujian ini dilakukan pemecah kalimat kedalam kata 1-grams, 3-grams, 4-grams, 5-grams dengan menerapakan teknik pembobotan *Binary Term*, TF-IDF dan *Term frequency(TF)*. Adapun sekenario pengujian bisa dilihat pada tabel 2.

Tabel 2. Model uji

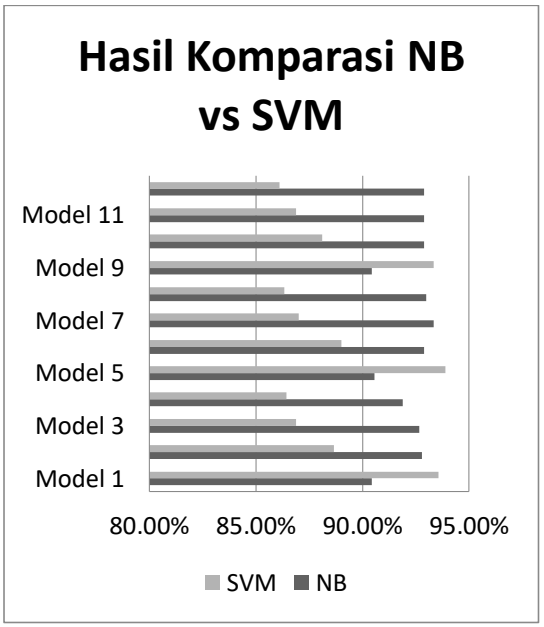
Model	Pembobotan	Token
Model 1	TF-IDF	1 Grams
Model 2		3 Grams
Model 3		4 Grams
Model 4		5 Grams
Model 5	TF	1 Grams
Model 6		3 Grams
Model 7		4 Grams
Model 8		5 Grams
Model 9	Binary Term	1 Grams
Model 10		3 Grams
Model 11		4 Grams
Model 12		5 Grams

Pada **tabel 2** adalah sekenario pengujian yang dilakukan, dimana terdapat 3 model yang digunakan. Setiap model mempunyai dan pemecah kalimat yang berbeda-beda. Sekenario pengujian model ini akan menjadi dasar perbandingan.

Model Uji menggambarkan model pengujian yang dibuat. Model ini dibuat untuk melihat hasil komparasi antara Naïve Bayes dan Support Vector Machine.

Selain untuk menbandingkan antara dua Machine Learning **table 2** menunjukan berbagai pembobotan dan juga pemisahan SMS, untuk melihat stabilitas dari Machine Learning.

Tabel 2 Model uji menggunakan 3 jenis pembobotan yaitu TF-IDF, TF dan Binary Term. Selain dibedakan dengan pembobotan juga dibedakan berdasarkan token yaitu: 1 grams, 3 grams, 4 grams dan yang terakhir 5 grams.

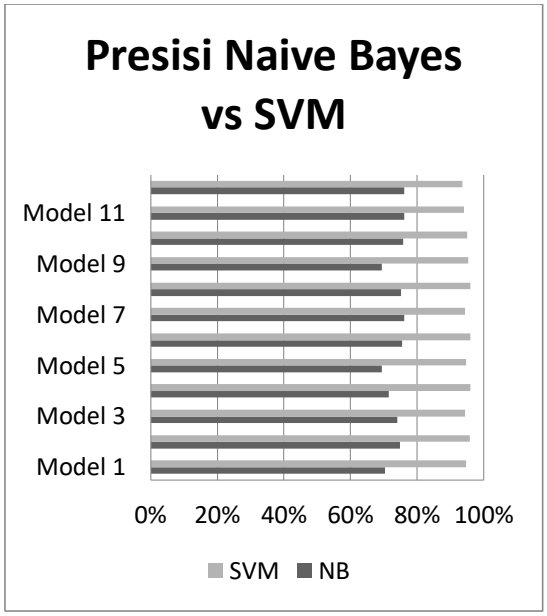


Gambar 2. Hasil akurasi dari model pengujian

Setelah dilakukan uji coba pada datasetnya maka diperoleh hasil akurasi yang diperlihatkan pada Gambar 2. Hasil Akurasi dari Model Pengujian.

Model 1 menunjukan hasil akurasi yang paling tinggi yaitu 93.56 % dengan menggunakan pembobotan TF-IDF dan token menggunakan 1-grams. TF-IDF menunjukan sebaran *term* yang

sering muncul dikalikan dengan *inverse*/kebalikan *term* dokumen yang jarang muncul ini artinya *term* memiliki nilai yang tengah –tengan pada kumpulan *term.* Sehingga dapat dikatakan tidak ada *term* yang memiliki nilai bobot nol, artinya semua data/sms dapat diklasifikasikan dengan baik oleh *Machine Learning* khususnya menggunakan SVM. Tetapi berbeda hasil akurasi yang ditunjukan oleh *Naïve Bayes* nilai akurasi yang paling rendah dibandingkan yang lain. Dari data yang didapat diketahui bahwa dengan data yang kurang bervariatif cocok menggunakan Naïve Bayes tetapi jika data lebih beragam disarankan menggunakan SVM.



Gambar 3. Perbadingan Presisi *Naïve Bayes* dan SVM

Naïve Bayes dapat mengklasifikasi dengan baik dan stabil tetapi untuk masalah presisi SVM menunjukan nilai yang baik. Presisi pada *Naïve Bayes* menunjukan hasil kurang baik dibandingkan dengan SVM, artinya masih banyak salah klasifikasi dataset jika menggunakan *Naïve Bayes*. Ini akan merugikan penerima SMS jika menggunakan *Naïve Bayes* sebagai *Machine Learning* pada ponsel, banyak Ham SMS yang masuk pada kategori Spam SMS begitupun sebaliknya. Hasil Presisi antara *Naïve Bayes* dan SVM ditunjukan pada Gambar 3. Perbandingan Persisi antara *Naïve Bayes* dan SVM.

V. KESIMPULAN DAN SARAN

Kesimpulan

Pada penelitian ini menunjukan bahwa *Naïve Bayes* memiliki hasil akurasi yang stabil dengan nilai rata-rata diatas 90%, tetapi memiliki selisih akurasi yang sedikit sekitar 3.37 % dibandingankan dengan SVM. Berbeda dengan dengan hasil presisi diantara *Naïve Bayes* dan SVM, SVM memiliki tingkat presisi yang baik dibandingkan dengan *Naïve Bayes*. SVM memiliki nilai rata-rata presisi sebesar 94,98 dan *Naïve Nayes* sebesar 73,69% itu artinya rata-rata selisih presisi antara keduanya yaitu sebesar 21,29%. Dari hasil perbandingan keduanya, SVM lebih unggul dibandingkan *Naïve Bayes*.

Saran

Untuk penelitian berikutnya bisa ditambahkan variable lain seperti lokasi sms spam, untuk mengetahui lokalisasi penerima spam sms yang tersebar di Indonesia, selain itu dicoba lakukan pembobotan selain menggunakan metode yang disebutkan sebelumnya menggunakan algoritma genetika, sebelum dilakukan kelasifikasi baik menggunakan SVM, maupun Naïve bayes.

DAFTAR PUSTAKA

Almeida TA, Gomes JM and Yamakami A (2011) *Contributions to the Study of SMS Spam Filtering* : New Collection and Results. , 1–4.

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

Apandi TH and Sugianto CA (2015) *Penyaringan Spam Short Message Service Menggunakan Support Vector Machine*. In: *Seminar Nasional Teknologi Informasi dan Komunikasi Terapan (SEMANTIK)*. 111–116.

Brown J, Shipman B and Vetter R (2007) SMS: *The Short Message Service*.

Brownlee J (2014) *How To Implement Naive Bayes From Scratch in Python*. [Online]

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

Hu XIA and Yan FU (2010) *Sampling Of Mass Sms Filtering Algorithm Based On Frequent Time-Domain Area*. , 548–551.

Jane S, Buckley M and Greene D (2012) *SMS spam filtering : Methods and data*. *Expert Systems With Applications*, 1–10. Available at: http://dx.doi.org/10.1016/j.eswa.2012.02.053.

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

Preoţiuc Pietro D and Florentina H (2012) *Unsupervised word sense disambiguation with N-gram features*. 41 (2012): . *Artificial Intelligence Review*. 41, 241–260.

Saini JR and Desai AA (2012) I*dentification of Most Frequently Occurring Lexis in Body-enhancement Medicinal Unsolicited Bulk e-mails Identification of Most Frequently Occurring Lexis in Body-enhancement Medicinal Unsolicited Bulk e-mails*.(May 2016).

Santoso B (2007) *Data Mining: Teknik Pemanfaatan Data untuk Keperluan Bisnis*. Yogyakarta: Graha Ilmu

Srivastava DK and Bhambhu L (2010) D*ata Classification Using Support Vector Machine*. *Journal of Theoretical and Applied Information Technology*. 12 (1), 1–7.