KLASIFIKASI EMAIL SPAM DENGAN METODE NAÏVE BAYES CLASSIFIER MENGGUNAKAN JAVA PROGRAMMING

Prasetyo Anugroho[1], Idris Winarno[2]., S.ST M.Kom, Nur Rosyid M[2]., S.Kom
[1] Mahasiswa, [2] Dosen Pembimbing
Politeknik Elektronika Negeri Surabaya
Institut Teknologi Sepuluh Nopember Kampus ITS Keputih Sukolilo Surabaya 60111, Indonesia
Tel:+62-31-7669770
Email: prara@student.eepis-its.edu

Abstrak

Internet telah menjadi salah satu hal yang penting dalam perkembangan sarana komunikasi. Salah satu fasilitas yang terdapat pada internet adalah surat elektronik atau yang lebih dikenal sebagai e-mail. Fasilitas e-mail yang mudah digunakan dan murah mengakibatkan banyaknya e-mail yang berisi iklan dan promosi bisnis masuk ke dalam inbox pengguna email. E-mail iklan inilah yang disebut sebagai spam mail. Untuk mencegah hal ini, dibuatlah software yang berguna sebagai spam filter untuk menyaring e-mail yang masuk ke dalam inbox pengguna fasilitas e-mail. Pemrograman spam filter pada tugas akhir ini menggunakan algoritma yang dinamakan Naive Bayes Classifier. Metode ini dipilih selain karena merupakan pengembangan terbaru dari pemrograman spam filter, juga karena algoritma ini memiliki tingkat keakuratan yang lebih tinggi dibandingkan dengan algoritma sebelumnya (contoh: NN Classifier).Dengan menggunakan Pemrograman Java, program spam filter dengan algoritma Naive Bayes Classifier telah berhasil dibuat. Dengan program ini dilakukan beberapa percobaan yang berhasil membuktikan bahwa algoritma ini mampu mengidentifikasikan spam, dengan beberapa syarat dan kondisi, secara lebih akurat.

Pembuatan program untuk sistem klasifikasi email dapat dilakukan 2 cara yaitu system klasifikasi dapat beroperasi pada mail client (offline) dan dapat beroperasi pada mail server (online).

***Kata Kunci**: spam, spam filter, naive bayes classifier, online, offline, mail server, mail client.*

1. PENDAHULUAN

1.1.LATAR BELAKANG

Perkembangan teknologi dewasa ini telah berkembang dengan pesat, khususnya dibidang internet. Dengan adanya internet, segala informasi dan berita dapat diterima dan diakses oleh setiap orang. Bahkan dengan internet, setiap orang dapat mengirim dan menerima pesan dari orang satu ke orang lainnya, yang lebih dikenal sebagai email. Tetapi tidak semua orang menggunakan fasilitas email dengan baik dan benar, bahkan dapat menyebabkan kerugian bagi orang lain. Hal ini dikarenakan fasilitas *e-mail* yang murah dan mudah digunakan oleh setiap orang, sehingga mengakibatkan banyaknya *spam- mail* yang masuk ke dalam *inbox* pengguna *e-mail*.

Spam-mail dapat didefinisikan sebagai "*unsolicited bulk e-mail*" yaitu *e-mail* yang dikirimkan kepada ribuan penerima (*recipient*). *Spam-mail* biasanya dikirimkan oleh suatu perusahaan untuk mengiklankan suatu produk, hal ini menyebabkan semakin padatnya queue atau antrian dari mail server yang telah digunakan. Pada survey yang dilakukan oleh Cranor & La Macchia (1998), ditemukan bahwa 10% dari mail yang diterima oleh suatu perusahaan adalah *spam-mail*. Tahun lalu, Spamcop, yang menjalankan servis untuk menerima laporan tentang spam, menerima lebih dari 183 juta laporan *spam*.[6]

Terlepas dari waktu yang terbuang untuk menghapus *spam-mail* dari *inbox*, *spam* juga merupakan pemborosan uang bagi *user* yang menggunakan koneksi *dial-up*. Selain itu *spam* juga membuang bandwidth dan dapat menyebabkan penerima di bawah umur mengakses situs-situs yang tidak seharusnya.

Pada proyek akhir ini, akan dibangun sebuah aplikasi mailclient yang salah satu fiturnya dapat menggolongkan atau mengklasifikasikan suatu email mana yang termasuk email spam, bukan spam ataupun email yang ragu-ragu antara spam atau bukan spam. Pemilihan terhadap metode naïve bayes pada sistem dalam melakukan klasifikasi, diharapkan sistem dapat mengklasifikasikan dengan tingkat keakuratan yang tinggi sehingga spam email dapat teridentifikasi dengan baik. Jika metode klasifikasi berhasil diterapkan pada mail client, maka selanjutnya akan diterapkan metode tersebut pada mail server.

1.2.RUMUSAN MASALAH

Berdasarkan uraian di atas, maka permasalahan yang timbul dalam pengerjaan proyek akhir ini antara lain adalah:

1. Bagaimana mendapatkan sebuah kata kunci yang mewakili *content* dari email?
2. Bagaimana membangun sistem yang dapat menghasilkan output email yang sudah terklasifikasi secara akurat dengan nilai error sekecil mungkin?
3. Bagaimana membangun sistem jaringan yang klasifikasi email secara otomatis pada *mail server*?

1.3.BATASAN MASALAH

Batasan masalah dalam pembuatan email client ini antara lain :

1. Mengklasifikasikan email yang **s**pam dan bukan (ham) pada email client yang telah dibangun.
2. Klasifikasi Spam email hanya memperhatikan header dan body dari email yang dalam bentuk bahasa inggris dan tidak dapat melakukan pengecekan terhadap sebuah attachment atau file.
3. Jika sistem klasifikasi email sudah dapat bekerja dengan baik maka akan diharapkan dapat bekerja pula pada mail server.

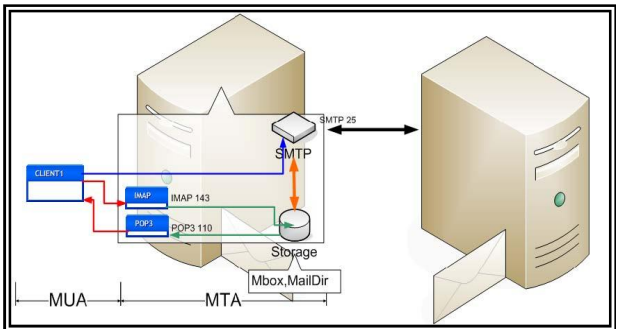
1.4.TUJUAN

Proyek akhir yang mempunyai ini bertujuan untuk membuat sebuah aplikasi yang dapat mengklasifikasikan email yang spam dan bukan spam(ham) secara otomatis dengan tingkat akurasi yang tinggi.

2. TEORI PENUNJANG

2.1.MAIL

Electronic-Mail (*E-Mail*) merupakan sebuah metode untuk mengirimkan pesan dalam bentuk digital. Pesan ini biasanya dikirimkan melalui medium internet. Sebuah pesan elektronis terdiri dari isi, alamat pengirim, dan alamat-alamat yang dituju. Sistem e-mail yang beroperasi di atas jaringan berbasis pada model *store and forward.* Sistem ini mengaplikasikan sebuah sistem server *e-mail* yang menerima, meneruskan, mengirimkan, serta menyimpan pesan-pesan user, dimana user hanya perlu untuk mengkoneksikan pc mereka ke dalam jaringan. *E-mail* dapat dianalogikan dengan kotak surat yang ada di kantor POS sedangkan server *e-mail* dapat diibaratkan sebagai kantor POS. Dengan analogi ini sebuah mail server dapat memiliki banyak account *e-mail* yang ada didalamnya. Penulisan e-mail dan *e-mail* sama saja. Namun lebih direkomendasikan untuk menuliskannya sebagai e-mail. Pada RFC, spelling e-mail yang digunakan adalah mail, dan sebuah e-mail dinamakan sebagai sebuah message. RFC yang baru dan grup IETF membutuhkan penulisan e-mail yang konsisten dari segi kapitalisasinya, penggunan underscorenya, serta ejaannya.



Gambar 1. Cara kerja *e-mail*

Cara kerja *e-mail* yang dapat dilihat pada Gambar 2.1 menunjukkan bahwa e-mail yang dikirim belum tentu akan diteruskan ke komputer penerima (*end user*), tapi disimpan/dikumpulkan dahulu dalam sebuah komputer server (*host*) yang akan online secara terus menerus (*continue*) dengan media penyimpanan (*storage*) yang relatif lebih besar dibanding komputer biasa. Hal ini bisa diibaratkan dengan sebuah kantor pos, jika sesorang mempunyai alamat *(mailbox)*, maka dia dapat memeriksa secara berkala jika dia mendapatkan surat. Komputer yang melayani penerimaan *e-mail* secara terus-menerus tersebut biasa disebut dengan *mailserver* atau *mailhost*.

2.2.SPAM MAIL

Pendefinisian spam *e-mail* berbeda-beda. Undang-undang CAN-SPAM memberikan definisi utama spam dengan menjelaskan apa yang (dan apa yang tidak) diperbolehkan bila mengirim *e-mail* komersial pemasaran. Undang-undang tersebut disahkan pada tahun 2004 oleh Federal Trade Commission, yang diperbarui tahun 2008. Selain FTC terdapat badan-badan lain yang mengklasifikasikan spam, yaitu Internet Service Provider (ISP). Internet Service Provider juga memiliki bagian besar dalam menentukan apa yang dianggap spam. ISP tidak mengandalkan CAN-SPAM sendirian untuk mendefinisikan spam karena di mata mereka spam '*didefinisikan oleh pengguna'*. Jika penerima *e-mail* mengelompokkan pesan *e-mail* sebagai spam dengan cara meletakkan di daftar pengirim yang diblokir mereka, menjatuhkannya di folder spam atau sekadar tidak konsisten membukanya, maka itu dianggap spam oleh ISP - terlepas dari apakah itu melekat pada masing-masing dan setiap CAN-SPAM aturan.

Berikut adalah tipe-tipe *e-mail* spam [2]:

a. Untuk Iklan: Spam dapat digunakan untuk mempromosikan suatu produk ataupun layanan, mulai dari produk software, perumahan real estate hingga produk kesehatan dan produk vitamin.
b. Untuk Mengirimkan Malware: Spam adalah salah satu cara utama untuk mendistribusikan virus dan malware. Dengan target yang bersifat individual, akan memperdaya korban untuk mempercayai bahwa mereka menerima dokumen penting atau file tertentu, yang sebenarnya mengandung malware.
c. Phishing: Bersembunyi dibalik nama-nama besar perusahaan besar, lembaga keuangan, lembaga pemerintah, lembaga amal, para phisher mencoba memikat korban untuk mengunjungi website palsu, dimana melalui website tersebut mereka dapat mencuri data keuangan pribadi atau informasi dengan mengenai identitas korbannya.
d. Scam: Mengirimkan *e-mail* sebagai pangeran dari Nigeria, pegawai bank dari Swiss, seorang anak kecil yang sakit keras, dan beberapa tipe lainnya, para scammer berusaha memperoleh simpati.
e. Pesan yang tak berarti: Sebuah potongan pesan sampah seperti ini dapat memenuhi inbox mail kita. Bahkan beberapa pesan seperti ini dapat mengkelabui teknologi spam filter, banyak pesan tak berarti ini dikirimkan tanpa tujuan yang jelas.

Perbedaan Spam dan Ham (bukan spam) berdasarkan struktur *e-mail* dapat diklasifikasikan sebagai berikut:

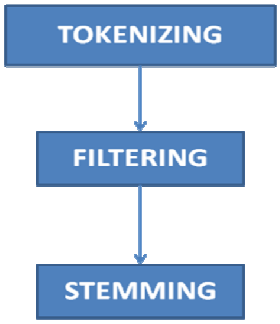
- Header
  Email header menunjukkan informasi perjalanan setiap email. Secara umum,

email header terdiri dari pengirim, jaringan dan penerima email [9].

- Subject
  Subject suatu *e-mail* biasanya merupakan suatu judul topic yang mewakili isi pada *e-mail*. Subject *e-mail* dapat dijumpai pada header setiap *e-mail*. Maka dapat dilihat pada gambar header spam *e-mail*, terdapat suatu kata "VIAGRA". Kata-kata tersebut sering dijumpai pada subject spam *e-mail*.
- Body
  Pada *e-mail*, body adalah isi dari suatu pesan *e-mail*, dan dengan adanya body *e-mail*, pengirim (sender) menyampaikan maksud yang akan disampaikan kepada penerima. Pada proyek akhir ini selain mengklasifikasikan *e-mail* spam dari header, dapat pula diklasifikasikan melalui bodynya. Karena dengan body *e-mail*, dapat ditentukan bahwa *e-mail* tersebut *e-mail* yang penting atau tidak.

2.3.TEXT MINING

Text mining mempunyai definisi sebagai menambang data yang berupa teks dimana sumber data biasanya didapat dari suatu dokumen dan tujuannya adalah mencari kata-kata yang dapat mewakili isi dari dokumen sehingga dapat dilakukan analisa keterhubungan antar dokumen.
Tahapan dari text mining :



Gambar 2. Bagan tahapan text mining

2.4.ALGORITMA NAÏVE BAYES

Bayesian filter atau Naïve Bayes Classifier merupakan metode terbaru yang digunakan untuk mengklasifikasikan sekumpulan dokumen. Algoritma ini memanfaatkan metode probabilitas dan statistik yang dikemukakan oleh ilmuwan Inggris Thomas Bayes, yaitu memprediksi probabilitas di masa depan berdasarkan pengalaman di masa sebelumnya. Dua kelompok peneliti, satu oleh Pantel dan Lin, dan yang lain oleh Microsoft Research memperkenalkan metode statistik Bayesian ini pada teknologi anti spam filter. Tetapi yang membuat algoritma Bayesian filtering ini popular adalah pendekatan yang dilakukan oleh Paul Graham [4].

Dasar dari teorema *naive* Bayes [4] yang digunakan dalam pemrograman adalah rumus Bayes berikut ini:

$$P(A|B) = (P(B|A) * P(A))/P(B) \ldots. (1)$$

Peluang kejadian A sebagai B ditentukan dari peluang B saat A, peluang A, dan peluang B. Pada pengaplikasiannya nanti rumus ini berubah menjadi :

$$P(Ci|D) = (P(D|Ci)*P(Ci)) / P(D) \ldots. (2)$$

Naïve Bayes Classifier atau bisa disebut sebagai multinomial naïve bayes merupakan model penyederhanaan dari algoritma bayes yang cocok dalam pengklasifikasian text atau dokumen [14].

Persamaannya adalah :

$$v_{MAP} = \arg\max P(v_j \mid a_1, a_2, \ldots. a_n) \ldots. (3)$$

menurut persamaan (3), maka persamaan (1) dapat ditulis

$$v_{MAP} = \arg\max_{v_j \in V} \frac{P(a_1 a_2 \ldots a_n \mid v_J) P(v_j)}{P(a_1 a_2 \ldots a_n)} \quad \ldots.(4)$$

$P(a_1, a_2, \ldots a_n)$ konstan, sehingga dapat dihilangkan menjadi

$$v_{MAP} = \arg\max_{V_j \in V} P(a_1 a_2 \ldots a_n \mid v_j) P(v_j) \quad \ldots (5)$$

Karena $P(a_1, a_2, \ldots a_n \mid v_j)$ sulit untuk dihitung, maka akan diasumsikan bahwa setiap kata pada dokumen tidak mempunyai keterkaitan.

$$v_{MAP} = \arg\max_{V_j \in V} P(v_j) \prod_i P(a_i \mid v_j) \quad \ldots (6)$$

Keterangan :

$$P(v_j) = \frac{|docs_j|}{|Contoh|} \quad \ldots..(7)$$

$$P(w_k \mid v_j) = \frac{n_k + 1}{n + |kosakata|} \quad \ldots.(8)$$

Dimana untuk :

$P(v_j)$: Probabilitas setiap dokumen terhadap sekumpulan dokumen.

$P(w_k|v_j)$: Probabilitas kemunculan kata w_k pada suatu dokumen dengan kategori class v_j

$|docs|$: frekuensi dokumen pada setiap kategori

$|Contoh|$: jumlah dokumen yang ada

n_k : frekuensi kata ke-k pada setiap kategori

kosakata : jumlah kata pada dokumen test

Pada persamaan (8) terdapat suatu penambahan 1 pada pembilang, hal ini dilakukan untuk mengantisipasi jika terdapat suatu kata pada dokumen uji yang tidak ada pada setiap dokumen data training.

Algoritma Naïve Bayes Classifier :

1.Learning (Pembelajaran)

Naïve Bayes adalah algoritma yang termasuk ke dalam supervised learning, maka akan dibutuhkan pengetahuan awal untuk dapat mengambil keputusan. Langkah-langkah :

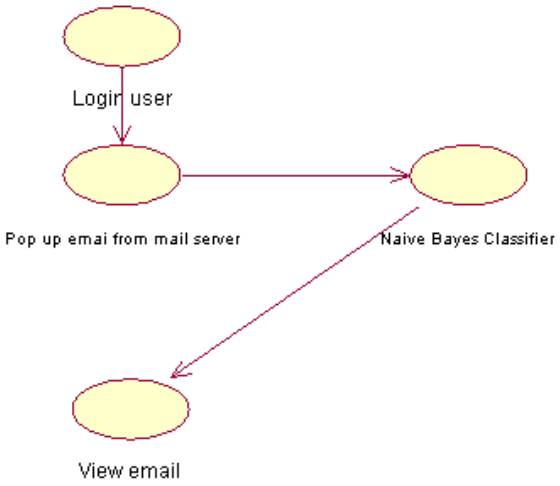
a. Step 1 : Bentuk vocabulary pada setiap dokumen data training
b. Step 2 : Hitung probabilitas pada setiap kategori $P(v_j)$
c. Step 3 : Tentukan frekuensi setiap kata w_k pada setiap kategori $P(w_k \mid v_j)$

2.Classify (Pengklasifikasian)

a. Step 1 : Hitung $P(v_j)\Pi P(a_i \mid v_j)$ untuk setiap kategori
b. Step 2 : Tentukan kategori dengan nilai $P(v_j)\Pi P(a_i \mid v_j)$ maksimal

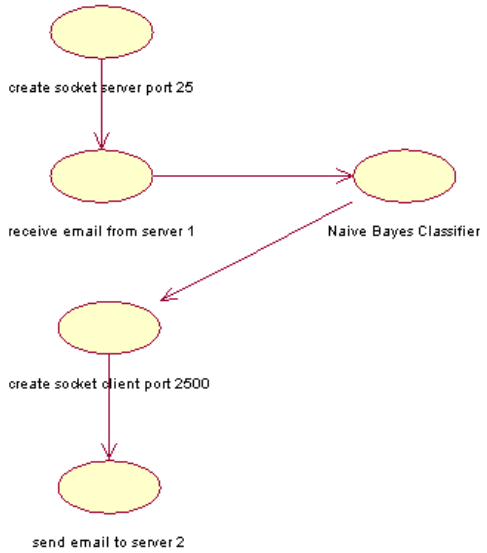
3. PERANCANGAN DAN IMPLEMENTASI

System klasifikasi pada mail client dibuat seperti halnya mail client pada umumnya, seperti Microsoft Office Outlook, Thunderbird dan lain-lain. Metode naïve bayes classifier yang telah dibangun sebelumnya, akan diletakkan pada program dari mail client itu sendiri. Seperti yang ditunjukan pada gambar 3.



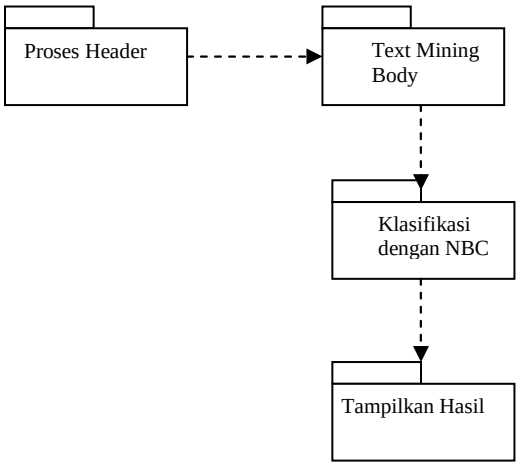
Gambar 3. Blog Diagram system offline

Jika system klasifikasi pada email pada mail client sudah dapat bekerja dengan baik, maka metode algoritma naïve bayes classifier akan diterapkan kedalam mail server atau bisa dikatakan dengan nama online. Dapat diilustrasikan seperti pada gambar 4.

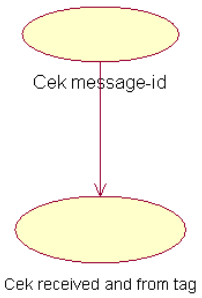


Gambar 4. Blog Diagram System Online

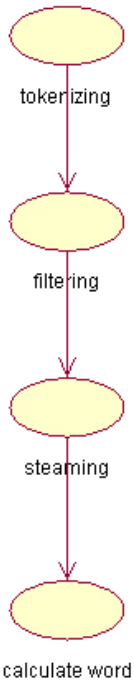
Use-Case Utama (*Architecturally Significant*) pada gambar 5 di atas adalah gambaran sistem secara garis besar yang dibedakan menjadi empat proses utama, yaitu proses header, proses *text mining*, proses pengklasifikasian dengan metode *naïve bayes classifier* dan bagaimana menampilkan hasil dari klasifikasi termasuk kategori spam atau regular email.



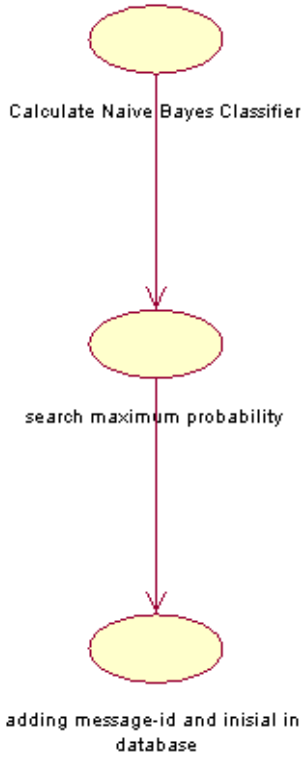
Gambar 5. Use Case Utama



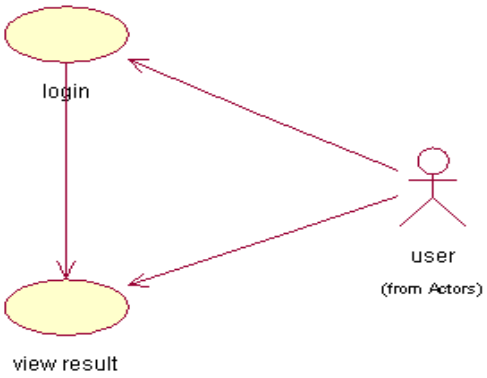
Gambar 6. Use Case Diagram Proses Header



Gambar 7. Use Case Diagram Text Mining



Gambar 8. Use Case Diagram Proses Naïve Bayes

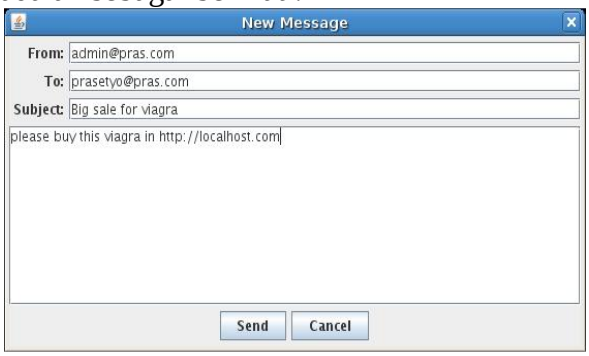


Gambar 9. Use Case Diagram Tampilkan hasil

4. UJI COBA DAN ANALISA

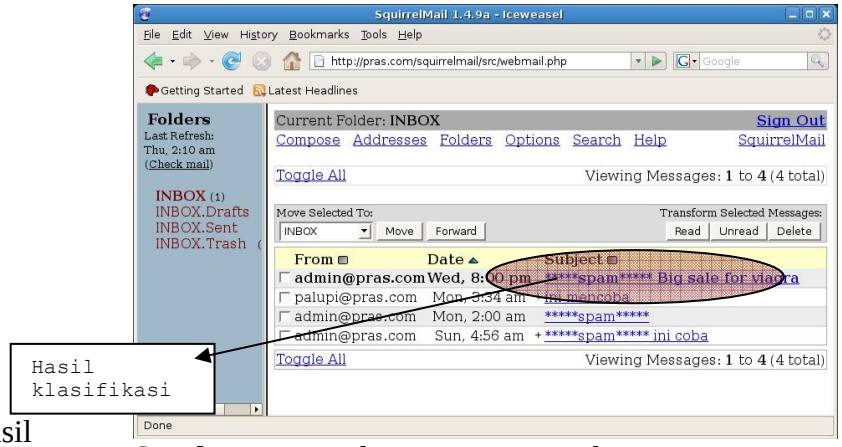
4.1.Percobaan 1

Pada percobaan ke-1 diilustrasikan dengan cara mengirimkan email yang mempunyai subject "Big sale for Viagra". Bagaimanakah hasil dari system klasifikasi dalam mengklasifikasikan tersebut. Hasil perhitungan probabilitas maksimum dari email, hasil klasifikasinya dan tampilannya adalah sebagai berikut :



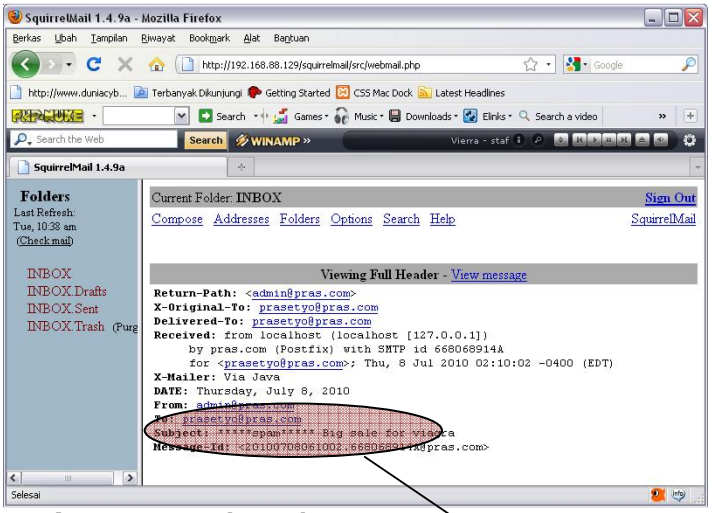
Gambar 10. Kirim email percobaan1

Pada gambar 10. merupakan suatu interface dari mail client pada system offline untuk mengirimkan email dengan isi subject yang mengandung kata "Viagra". Dan ternyata hasil yang didapat setelah mengalami proses klasifikasi adalah email tersebut merupakan SPAM. Hal ini dapat dibuktikan pada gambar 11.



Gambar 11. Hasil pengiriman email percobaan 1

Hal ini dapat juga dilihat pada gambar 12 yaitu header email, bahwa header email telah mengalami perubahan. Jika email tersebut telah terklasifikasi sebagai spam, maka subject pada header email akan ditambahi dengan kata "****spam****".



Gambar 12. Isi email percobaan 1

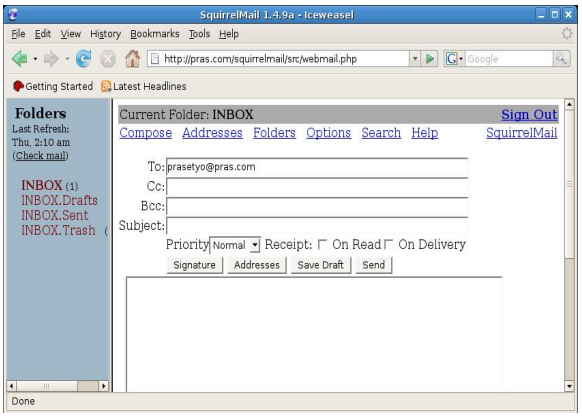
Analisa Percobaan 1

Dari hasil yang telah ditampilkan pada percobaan 1 diatas dapat diketahui bahwa email tersebut telah diklasifikasikan sebagai SPAM. Menurut algoritma yang telah dioperasikan pada proyek akhir ini bahwa sebelum email memasuki program metode naïve bayes, subject email tersebut dicek terlebih dahulu. Jika mengandung kata "Viagra" maka dapat dipastikan bahwa email adalah spam dan tidak akan menjalani proses

klasifikasi dengan metode naïve bayes. Maka dengan begitu perhitungan probabilitas pada email tidak perlu dilakukan. Dengan segera subject pada header email akan diubah.

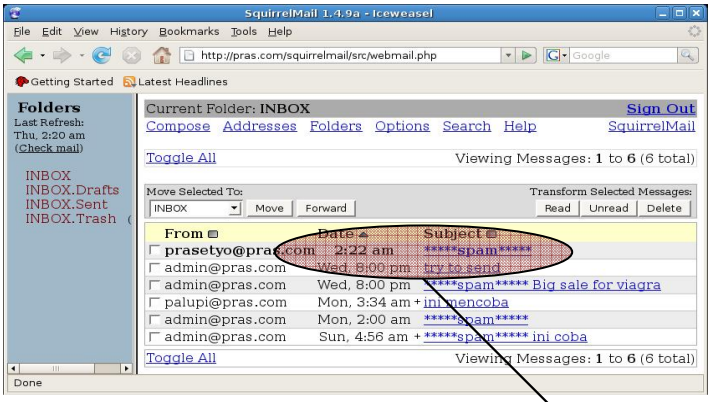
4.2.Percobaan 2

Dan pada percobaan 3 akan dicoba mengirimkan email dengan subject yang kosong dan email body yang kosong juga. Hal ini kadang dilakukan seorang user sewaktu user tersebut lupa dalam mengisikan bagian-bagian email yang akan dikirim. Dengan adanya hal seperti demikian,maka dikhawatirkan program tidak dapat memfilter email dengan format yang kosong seperti pada gambar 13. Untuk email pada percobaan 3 ini dikirim melalui squirrelmail karena pada mail client pada system offline tidak akan dapat mengirim email dalam bentuk kosongan.



Gambar 13 Kirim email percobaan 3

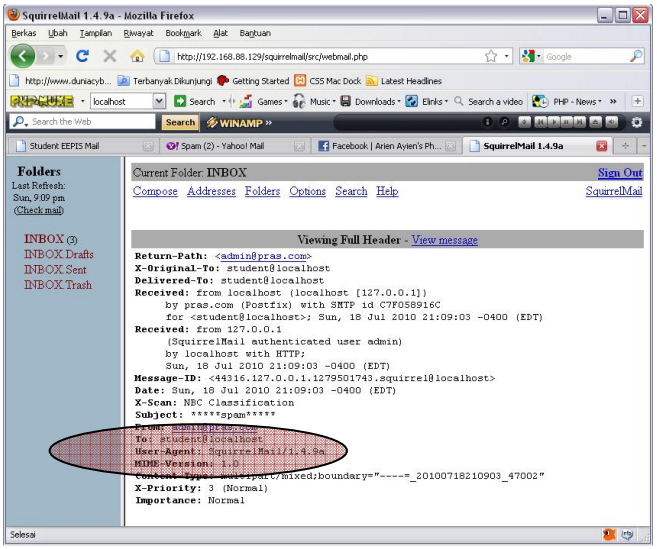
Setelah dikirim, ternyata email telah dianggap sebagai spam, hal ini dapat terlihat pada gambar 14. Dan berhubung subject email tidak terisi, maka secara langsung email subject email bertuliskan "****spam****".



Gambar 14 Hasil pengiriman email percobaan 3

Hasil klasifikasi

Dengan dinyatakannya email sebagai SPAM, maka header pada email tersebut juga mengalami perubahan. Perubahan header ini bersifat permanen, dan jika terdapat mail client yang melakukan pop up dari server maka akan subject sudah diubah.



Gambar 15. Header email percobaan 3

Analisa Percobaan 3

Setelah email dikirim, email akan masuk terlebih dahulu pada system klasifikasi secara online, dan yang pertama perlu diperiksa adalah subject email. Berhubung subject email yang dikirim berupa kosongan, maka system akan mengecek isi dari email, apakah dalam format bahasa Indonesia atau bahasa inggris. Tetapi seperti pada gambar 13, isi dari email kosong, maka menurut perhitungan :

Jumlah kata pada database bukan spam(nk) =5710
Jumlah kata pada database spam(nk) =5531
P(spam)=60/120 = 0.5
P(bukan spam)=60/120 = 0.5
Karena isi dari email adalah kosong, maka
P(w|spam)= 0.5
P(w|bukan spam)= 0.5
P(w|bukan spam) > P(w|spam) maka hasilnya adalah SPAM

4.3.Analisa secara umum

Berikut merupakan suatu percobaan secara umum untuk perhitungan nilai error dengan jumlah data training yang berbeda beda dan table 1 merupakan data-datanya

Jum Data	Jum Spam	Jumlah Bukan Spam	Jumlah Keyword pada Spam	Jumlah Keyword pada Bukan Spam	Nilai error
20	10	10	1132	914	1
40	20	20	4780	1587	12
60	30	30	5112	4067	3
80	40	40	5378	5043	3
93	47	46	5513	5710	3
120	60	60	7964	8227	3

Tabel 1. Data training dan tingkat error

Dan table 2 merupakan hasil dari klasifikasi email secara terperinci untuk 120 data training yang telah diklasifikasi. Terdapat 3 kesalahan (error) yang telah dihasilkan yaitu 2 data yang merupakan spam, diklasifikasi sebagai bukan spam dan 1 data bukan spam diklasifikasi sebagai spam. Maka prosentase kesalahannya mencapai,

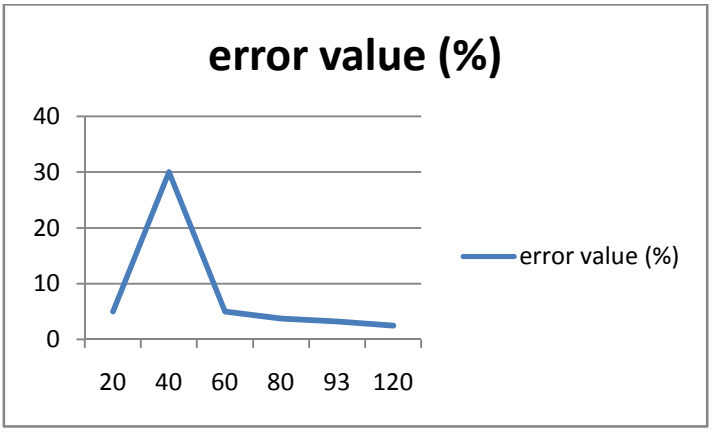
*error=3/120 * 100% = 2.5 %*

Dengan banyaknya data training dengan tingkat kesalahan error seperti demikian, maka 120 data training tersebut akan dijadikan sebagai database sebagai keyword untuk menyeleksi email data uji pada proyek akhir ini. Dan diharapkan dapat mengecilkan tingkat error dalam menyeleksi atau mengklasifikasi email (data uji).

Email yang Masuk	Hasil Klasifikasi Email	Nilai Kebenaran
Spam 1	SPAM	Benar
Spam 2	BUKAN SPAM	Salah
Spam 3	SPAM	Benar
Spam 4	SPAM	Benar
Spam 5	SPAM	Benar
Spam 6	SPAM	Benar
Spam 7	SPAM	Benar
Spam 8	SPAM	Benar
Spam 9	SPAM	Benar
Spam 10	SPAM	Benar
Spam 11	SPAM	Benar
Spam 12	SPAM	Benar
Spam 13	SPAM	Benar
Spam 14	SPAM	Benar
Spam 15	SPAM	Benar
Spam 16	SPAM	Benar
Spam 17	SPAM	Benar
Spam 18	SPAM	Benar
Spam 19	SPAM	Benar
Spam 20	SPAM	Benar
Spam 21	SPAM	Benar
Spam 22	SPAM	Benar
Spam 23	SPAM	Benar
Spam 24	SPAM	Benar
Spam 25	SPAM	Benar
Spam 26	SPAM	Benar
Spam 27	SPAM	Benar
Spam 28	SPAM	Benar
Spam 29	SPAM	Benar
Spam 30	SPAM	Benar
Spam 31	SPAM	Benar
Spam 32	SPAM	Benar
Spam 33	BUKAN SPAM	Salah
Spam 34	SPAM	Benar
Spam 35	SPAM	Benar
Spam 36	SPAM	Benar
Spam 37	SPAM	Benar
Spam 38	SPAM	Benar
Spam 39	SPAM	Benar
Spam 40	SPAM	Benar
Spam 41	SPAM	Benar
Spam 42	SPAM	Benar
Spam 43	SPAM	Benar
Spam 44	SPAM	Benar
Spam 45	SPAM	Benar
Spam 46	SPAM	Benar
Spam 47	SPAM	Benar
Bukan Spam 1	BUKAN SPAM	Benar

Bukan Spam 2	BUKAN SPAM	Benar
Bukan Spam 3	BUKAN SPAM	Benar
Bukan Spam 4	BUKAN SPAM	Benar
Bukan Spam 5	BUKAN SPAM	Benar
Bukan Spam 6	BUKAN SPAM	Benar
Bukan Spam 7	BUKAN SPAM	Benar
Bukan Spam 8	BUKAN SPAM	Benar
Bukan Spam 9	BUKAN SPAM	Benar
Bukan Spam 10	BUKAN SPAM	Benar
Bukan Spam 11	BUKAN SPAM	Benar
Bukan Spam 12	BUKAN SPAM	Benar
Bukan Spam 13	BUKAN SPAM	Benar
Bukan Spam 14	BUKAN SPAM	Benar
Bukan Spam 15	BUKAN SPAM	Benar
Bukan Spam 16	BUKAN SPAM	Benar
Bukan Spam 17	BUKAN SPAM	Benar
Bukan Spam 18	BUKAN SPAM	Benar
Bukan Spam 19	BUKAN SPAM	Benar
Bukan Spam 20	BUKAN SPAM	Benar
Bukan Spam 21	BUKAN SPAM	Benar
Bukan Spam 22	BUKAN SPAM	Benar
Bukan Spam 23	BUKAN SPAM	Benar
Bukan Spam 24	BUKAN SPAM	Benar
Bukan Spam 25	BUKAN SPAM	Benar
Bukan Spam 26	BUKAN SPAM	Benar
Bukan Spam 27	BUKAN SPAM	Benar
Bukan Spam 28	BUKAN SPAM	Benar
Bukan Spam 29	BUKAN SPAM	Benar
Bukan Spam 30	BUKAN SPAM	Benar
Bukan Spam 31	BUKAN SPAM	Benar
Bukan Spam 32	BUKAN SPAM	Benar
Bukan Spam 33	BUKAN SPAM	Benar
Bukan Spam 34	BUKAN SPAM	Benar
Bukan Spam 35	BUKAN SPAM	Benar
Bukan Spam 36	BUKAN SPAM	Benar
Bukan Spam 37	SPAM	Salah
Bukan Spam 38	BUKAN SPAM	Benar
Bukan Spam 39	BUKAN SPAM	Benar
Bukan Spam 40	BUKAN SPAM	Benar
Bukan Spam 41	BUKAN SPAM	Benar
Bukan Spam 42	BUKAN SPAM	Benar
Bukan Spam 43	BUKAN SPAM	Benar
Bukan Spam 44	BUKAN SPAM	Benar
Bukan Spam 45	BUKAN SPAM	Benar
Bukan Spam 46	BUKAN SPAM	Benar

Tabel 2. Hasil Klasifikasi 93 Data training



Gambar 16. Grafik dari tingkat error

Menurut yang telah ditampilkan pada table 4.1 dan grafik pada gambar 16 bahwa tingkat error lebih banyak ditunjukkan pada percobaan yang kedua disaat data training yang dipakai mencapai 40. Karena dapat dilihat pada kolom keyword untuk database spam dan yang bukan spam, selisih dari banyaknya keyword pada kedua kategori terlalu banyak. Sehingga memicu tingkat error yang lebih besar dibandingkan yang lainnya. Ini merupakan suatu kelemahan dari metode naïve bayes classifier dalam mengklasifikasikan data berupa string. Karena metode tersebut membutuhkan selisih dari jumlah keyword pada database data training yang tidak terlalu banyak.

5. PENUTUP

5.1.KESIMPULAN

Dari hasil percobaan dan analisa yang dilakukan maka dapat diambil kesimpulan:

1. Klasifikasi email spam dengan metode naïve bayes classifier yang dilakukan secara offline dan online tidak memiliki perbedaan dalam mengklasifikasikan email spam secara tepat dengan tingkat error yang kecil.
2. System klasifikasi email spam dengan metode naïve bayes classifier mempunyai tingkat error yang besar jika terdapat selisih pada jumlah keyword yang ada di data training.

5.2.SARAN

1. Karena pendataan kata untuk proses stoplist masih dilakukan secara manual, maka untuk pengembangan program ini selanjutnya diharapkan daftar kata dalam *stoplist* dapat dibakukan agar hasil text miningnya lebih optimal.
2. Pada proses steming dengan metode porter masih terdapat kesalahan dalam pembakuan kata yang berbahasa inggris dan diharapkan dapat dibakukan lagi.
3. Proses klasifikasi pada proyek akhir ini dapat diintegrasi dengan software lain dalam hal mengklasifikasikan email misalnya clamav.
4. Pada system online dari proyek akhir ini terdapat kelemahan metode dalam menangkap email dari mail server yang mengirim dan membutuhkan waktu yang cukup lama. Oleh karena itu dibutuhkan sebuah metode atau pola dalam menangkap email yang tepat dan mengklasifikasikannya.

6. DAFTAR PUSTAKA

[1] Website:http://www.emailaddressmanager.com/tips/email.html

[2] Website:http://lecturer.eepis-its.edu/~iwanarif/kuliah/dm/ 6Text%20Mining.pdf

[3] Roderick, W, Smith.2002. *Pull Mail Protocols : IMAP and POP3*. 257-282

[4] Rachli, Muhamad.2007. *Email Filtering menggunakan Naïve Bayesian*. Bandung : Tugas Akhir Jurusan Teknik Elektro Institut Teknologi Bandung

[5] Tapen, Panji. 2008. *Email Spam Filtering*. http://panjitapen.wordpress.com/2008/01/27/email-spam-filtering [12 Januari 2010]

[6] Arief .2004. *Spam: Dampak dan Resikonya.* http://www.ebizzasia.com [12 Januari 2010]

[7] Cakrawala .2007. *E-mail, Kendala dan Permasalahannya*. http://www.tnial.mil.id [17 Juli 2009]

[8] Magdalena, Merry. 2008. *Mengapa Alamat E-mail Kita Dicintai Spam*. http://www.netsains.com [17 Juli 2009]

[9] -------. 2007. *How to Read Email Headers.* http://www.emailaddressmanager.com/tips/email.html [17 Juli 2009]

[10] Spykerman, Mike. 2007. *How to Effectively Stop Spam and Junk Mail: Identifying The Most Common Spam Characteristics*. http://www.policypatrol.com

[11] -------. 2007. *Email Headers Comparison: Spam versus Regular Mail*. http://www.emailaddressmanager.com/tips/email.html

[12] -------. 2002. *The English Porter Stemmning Algorithm*. http://snowball.tartarus.org [

[13] Geeta, Gayathri Ravichandran. 2007. *Text Classification: An Application of Naïve Bayes Classifier*. Departement of Computer Science SUNY Stony Brook

[14] JavaMail™ API Design Specification. Sun Microsystems USA, Desember 2005