PENERAPAN METODE PEMBEDA MARKOV PADA PROSES PEMFILTERAN EMAIL SPAM

Sayed Fachrurrazi, S.Si., M.Kom
Program Studi Teknik Informatika, Universitas Malikussaleh
Reuleut, Aceh Utara, Aceh-Indonesia

E-mail: sayedfachrurrazi@gmail.com

ABSTRAK

Aplikasi spam filtering ini dibangun dengan menggunakan bahasa pemrograman Microsoft Visual Basic 6.0. Metode pembeda markov digunakan dalam melakukan filterisasi email yang diterima. Pembuatan fitur menggunakan Sparse Binary Polynomial Hash (SBPH) dengan skema pembobotan Exponential Super increasing Model (ESM). Metode pembeda markov mengklasifikaskan email menjadi email spam dan legitimate secara otomatis serta mengurangi kesalahan klasifikasi email legitimate menjadi email spam. Pada penelitian ini untuk mendapatkan tingkat akurasi pengklasifikasian email menjadi spam dan legitimate maka training data merupakan email dari account yang sama dengan email account yang akan difilter. Dari hasil penelitian ini yang didapat 69% tingkat keakuratannya.

Kata kunci: *Email, Spam, Filter Spam, Legitimate, Metode, Pembeda Markov, SBPH*

PENDAHULUAN

Pertumbuhan yang cepat dari internet, dalam hal ini komunikasi lewat electronic mail (email) menjadi salah satu bentuk komunikasi yang paling cepat ekonomi. Sebuah pesan email yang dikirim kepada sejumlah besar orang tampa peretujuan dari orang terebut, biasa disebut sebagai unsolicited

commercial email email (UCE), spam email, junk mal, bulk mail atau email sampah. Masalah email sampah (spam atau junk email) merupakan salah satu masalah yang dihadapi pada dunia internet. Untuk menyeleksi email yang datang secara manual akan membutuhkan waktu yang sangat banyak. Serta akan memakan kapasitas penyimpanan email yang akan memenuhi tempat penyipanan email-email tersebut. Sapam adalah pengguna perangkat elektronik utuk mengirimkan pesan secara bertubi-tubi tampa dikehendaki oleh penerimanya. Orang yang melakukan spam disebut spammer. Tindakan spam dikenal dengan nama spamming. Bentuk spam yang dikenal secara umum meliputi : spam surat elektronik, spam pesan instan, spam usernet news grop, spam mesin pencari informasi web (web search engine spam), spam blog, spam wiki, spam iklan baris daring, spam jejaring sosial. Beberapa contoh lain dari spam, yaitu ponsel berisi iklan, surat masa singkat (SMS) pada telepon genggam, berita dalam suatu forum kelompok warta berisi promosi barang yang tidak terkait dengan kegiatan kelompok warta tersebut, spamdexing yang menguasai suate mesin pencari (search engine) untuk mencari popularitas bagi suatu URL tertentu, berita yang tak berguna dan masuk dalam blog, buku tamu situs web, spam transmisi faks, iklan televisi dan spam jaringan berbagi. Spam dikirimkan oleh pengiklan dengan biaya operasional yang sangat rendah, karena spam tidak memerlukan senarai (mailing list) untuk mencapai para pelanggan-pelanggan yang diinginkan. Karena hambatan masuk yang rendah, maka banyak spammers yang muncul dan jumlah pesan yang tidak diminta menjadi sangat tinggi. Akibatnya, banyak pihak yang dirugikan. Selain pennguna Internet itu sendiri, ISP (Penyelenggara Jasa Internet atau Internet Service Provider), dan masyarakat umum juga merasa tidak nyaman. Spam sering mengganggu dan terkadang menipu penerimanya. Berita spam termasuk dalam kegiatan melanggar hukum dan merupakan perbuatan pidana yang bisa ditindak melalui undang-undang Internet. Spam memang menjengkelkan dan sangatmerugikan, bayangkan saja ibarat tamu tak diundang, mereka masuk ke rumah kita dengan ngomong seenaknya sendiri tampa memperhatikan etiet dan tata cara yang ada. Untuk itu kita harus aktif untuk tidak memiarkan spam berkeliaran di inbox email. Berdasarkan latar belakang masalah peneliti tertarik untuk meneliti tentang "Pemfilteran Email Spam dengan Menggunakan Metode Pembeda Markov".

DASAR TEORI

a. Email Spam

Spam-mail dapat didefinisikan sebagai "unsolicited bulk e-mail" yaitu email yang dikirimkan kepada ribuan penerima (recipient). Spam mail biasanya dikirimkan oleh suatu perusahaan untuk mengiklankan suatu produk. Karena fasilitas e-mail yang murah dan kemudahan untuk mengirimkan ke berapapun jumlah penerima, maka spam mail menjadi semakin merajalela. 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 (www.spamcop.net), yang menjalankan servis untuk menerima laporan tentang spam, menerima lebih dari 183 juta laporan spam.

b. Spam Filter

Spam filter dapat diartikan juga sebagai software anti-spam. Software ini menganalisa email yang datang dan menggunakan sejumlah metode untuk menentukan apakah email yang diterima sah atau tidak. Jadi jawaban untuk apakah spam filter bekerja adalah ya. Namun seberapa jauh keberhasilannya adalah masalah yang lain lagi. Ini ditentukan oleh spam filter yang anda miliki, spam filteryang satu lebih baik daripada yang lainnya.

c. Dampak buruk SPAM

SPAM mudah dilakukan dikarenakan Spammers selain umumnya menggunakan mail server orang lain, juga alamat e-mail aspal (asli tapi palsu); alamat e-mail tersebut memang benar ada tapi si pengirimnya bukan yang punya. Mengirim e-mail menggunakan alamat e-mail aspal sangat dimungkinkan karena protokol SMTP (Simple Mail Transfer Protocol) yang digunakan dalam pertukaran e-mail tidak pernah memverifikasi alamat e-mail dengan alamat IP-nya. Artinya, orang bebas mengirim e-mail dari manapun (dari alamat IP apapun) dengan menggunakan alamat e-mail siapapun.

d. Cara Mengurangi SPAM

Gunakan fasilitas mail filtering yang ada di Outlook Express dan nestcape messenger, kemudian buat rule supaya semua mail dengan isi spam, atau dari alamat tertentu yang biasanya mengirim spam di delete langsung dari server tanpa perlu di download sama sekali. Pada Outlook Express, tandai dulu salah satu mailnya, setelah itu pilih ’Message à Block àà Message Rules à Blocked Sender List’. Pada Netscape Messenger, fasilitas ini diakses melalui ’Edit à Message Filter’. Maka anda dapat langsung menghapus e-mail yang tak diinginkan tersebut.

METODOLOGI

Penelitian ini dengan mengumpulkan dan mempelajari literatur yang berkaitan dengan *Email Spam*, dengan menggunakan metode pembeda *Markov*. Sumber literatur berupa buku teks, paper, jurnal, karya ilmiah, dan situs-situs penunjang lainnya.

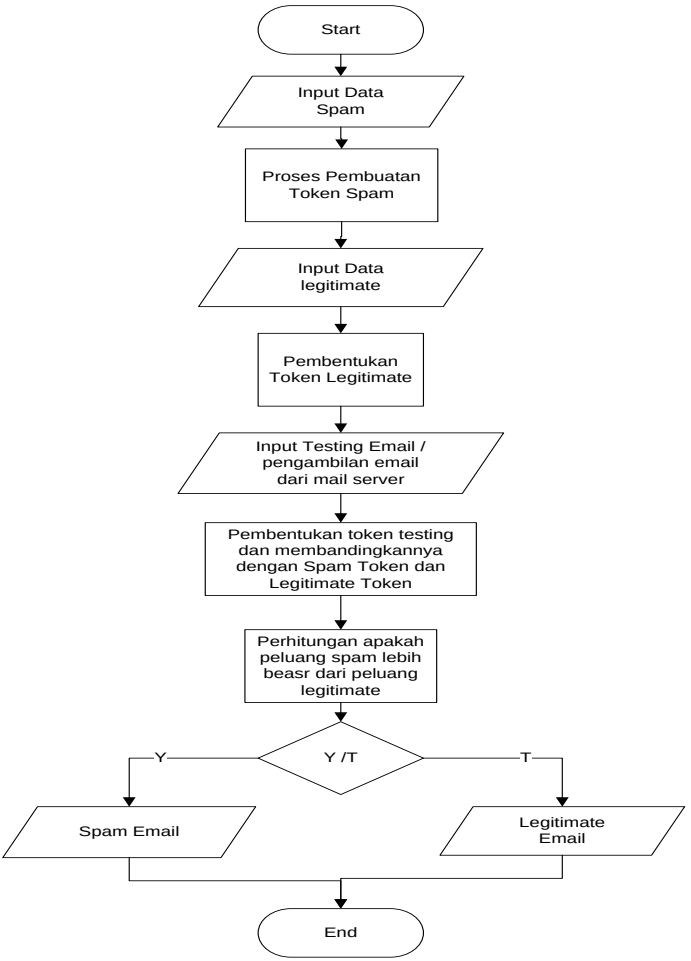
1.1. Alat Penelitian dan Bahan

Pada penelitian ini alat penelitian yang digunakan berupa perangkat keras dan perangkat lunak sebagai berikut:

a. Perangkat keras (hardware)
   1. Pentium (R) Dual-Core CPU T4200@2,0 Ghz
   2. Ram 1GB
   3. Hardisk 250GB
   4. Keybord, mouse

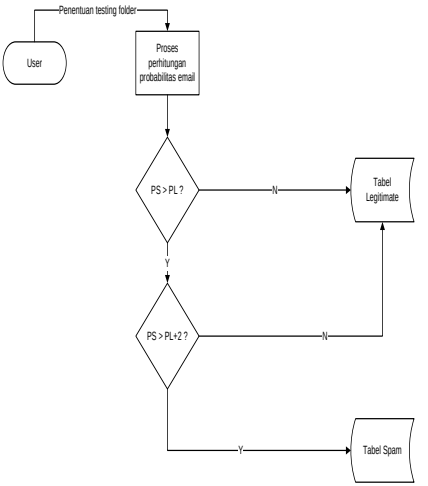
b. Perangkat lunak (software)
   1. Sistem Operasi Windows XP
   2. Visual Basic 6,0 sebagai bahasa pemograman
   3. MySQL untuk *database server*.
   4. Microsoft Office Word 2007
   5. Microsoft Office Visio 2007

Adapun Flowchart sistem ini dirancang untuk mengetahui langkah-langkah proses dalam sistem yang akan dibangun:



Gambar 3.1. Flowhart sistem

Dari gambar di atas maka dapat digambarkan schema sistem dari apliksi penfilteran email spam adalah sebagai beriut :



Gambar 3.2 Data Flow Diagram Pemfilteran Email

1.2. Proses Penelitian

Proses klasifikasi data yang akan digunakan dalam membangun aplikasi ini yaitu terbagi dua tahap antara lain :

a. Proses pembuatan data training

Data training adalah data yang akan digunakan dalam system untuk pembuatan token dimana token-token tersebut mencerminkan ciri-ciri atau kebiasaan dari data yang ditraining dalam hal ini yang dicari adalah cirri-ciri atau kebiasaan dari email spam dan email nonspam.

b. Proses pembuatan data testing

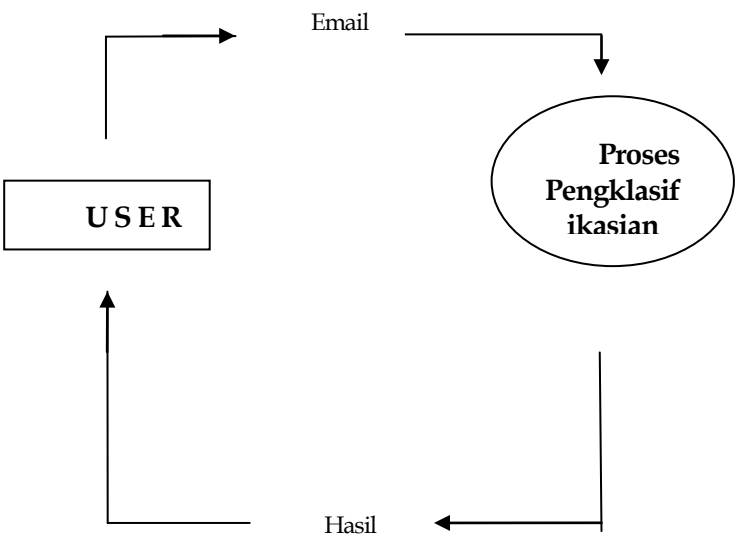
Data testing adalah data yang akan difilter oleh aplikasi ini, dimana data testing merupakan sebuah folder testing yang berisiskan file email berektensi .msg yang belum diketahui apakah email tersebut merupakan email spam atau bukan spam.

PERANCANGAN SISTEM

Pada tahap ini pengunaan notasi sangat membantu sekali dalam komunikasi dengan pemakai sistem, secara logika diagram yang menggunakan notasi ini biasanya dipakai untuk mengambarkan Diagram

Konteks dan Diagram Arus Data (DAD). Perancangan sistem merupakan gambaran atau sketsa dari alur proses sistem pengolahan data. Rancangan suatu sistem dapat menggunakan Diagram Arus Data (DAD) atau *Data Flow Diagram* (DFD).

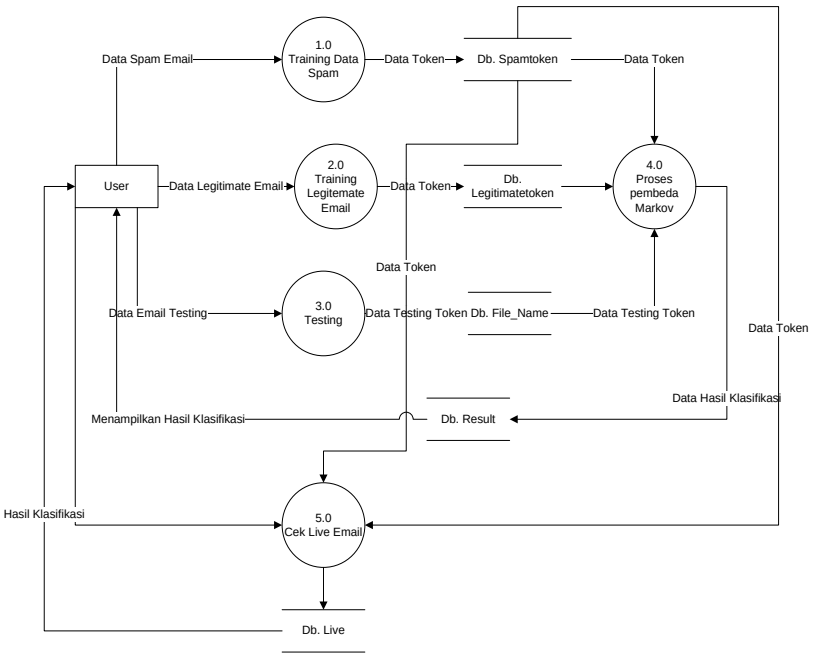
a. Menggunakan diagram konteks (*Contexs Diagram*) atau hubungan antara masing-masing komponen sistem yang terkait.
b. Menggunakan DFD (*Data Flow Diagram*) sistem yang merupakan penjelasan lebih detail lagi dari diagram konteks sistem tersebut.
c. Menggambarkan desain database (desain tabel), relasi antar table dan Interface input dan Output sistem secara umum.



Gambar 4.1. ***Diagram Konteks***

Keterangan :

a. User melakukan penentuan lokasi email spam folder, legitimate folder, testing folder.
b. Sistem melakukan pengklasifikasian.
c. User mendapatkan hasil klasifikasi.



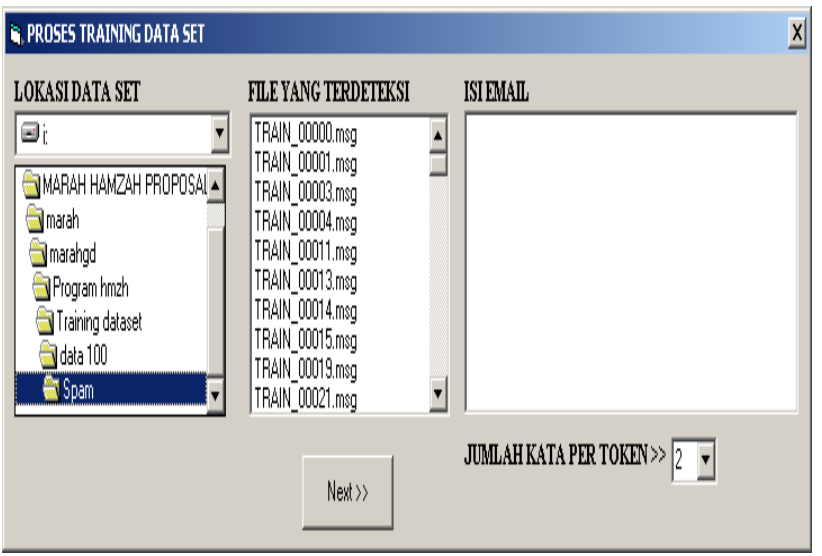
Gambar 4.2 Data Flow Diagram Leve 0

Penjelasan dari gambar DFD sistem diatas adalah sebagai berikut :

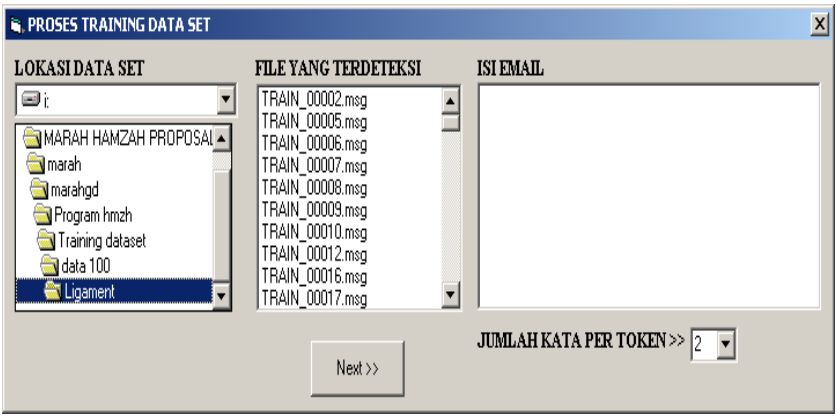
a. Training Data Spam adalah proses pembentukan token dari file-file yang telah ditentukan kemudian token-token tersebut disimpan dalam database Data Spam Token
b. Training Legimate Email adalah proses pembentukan token dari file-file yang telah ditentukan dan disimpan dalam database Legimate Token.
c. Testing adalah proses pengambilan token dari file testing email yang telah ditentukan kemudian disimpan database File Name.
d. Proses pembeda Markov adalah proses membandingkan token testing terhadap masing-masing token spam dan legimate yang hasil prosesnya disimpan dalam database Hasil.
e. Cek Live Email adalah proses pengecekan email langsung dari account pengguna dan hasil proses disimpan dalam database Hasil Live.

TAMPILAN APLIKASI

Gambar 4.4 menu utama



Gambar 4.5 Proses Training Spam Dataset



Gambar 4.6 Proses Training Legitimemate Dataset

Token
Re: How
to manage
multiple Internet
connections?
linux.ie mailing
list memberships
reminder
Re: Apple
Sauced...again
Re: results
for giant
mass-check (phew)
%post, %postun
etc
Re: Fwd:
Re: Kde
3.5 ...
Re: (no

Gambar 4.7 token nonspam terbentuk

Token
One of
a kind
Money maker!
Try it
for free!
link to
my webcam
you wanted
[SPAM] Give
her 3
hour rodeo
Best Price
on the
netf5f8ml
Enter now,
hibody, 75%
off
[SPAM] Summer

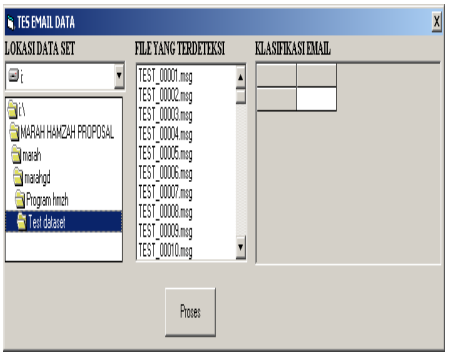
Gambar 4.8 token spam terbentuk

HASIL UJI COBA

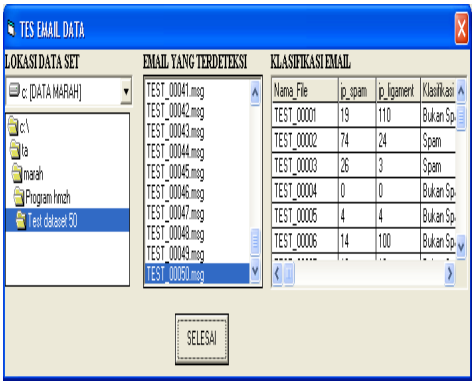
Peneliti melakukan beberapa langkah yang harus dilakukan dalam memfilter email antara lain :

a. Pada menu utama pilihlah Cek Email Data
b. Maka akan muncul form pengecekan email dataset kemudian tentukan test dataset folder maka file email yang akan difilter akan muncul ada File yang terdeksi
c. Kemudian klik proses untuk memulai pemfilteran email, maka pengklasifikasikan email akan ditampilkan

Berikut adalah proses tes email data.

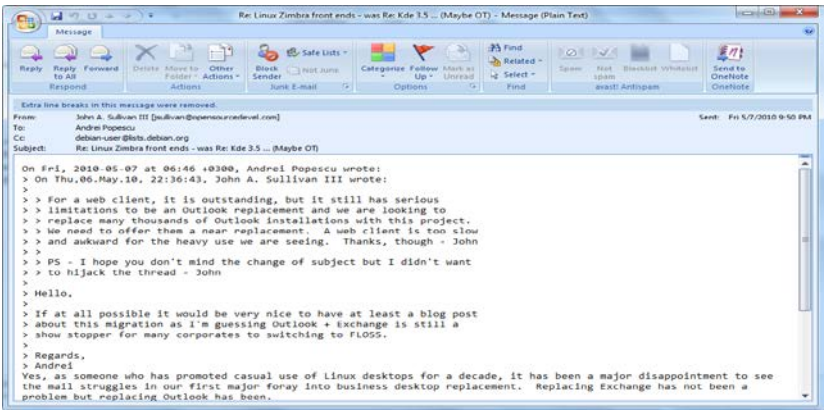


Gambar 4.9 Form Pemfilteran Email



Gambar 4.10 Hasil klasifikai email

Untuk mengecek kebenaran sistem maka kita akan menghitung secara manual sebuah email yang terklasifikasi sebagai spam atau legitimate.



Gambar 4.11 Data Flow Diagram Fungsi Markov

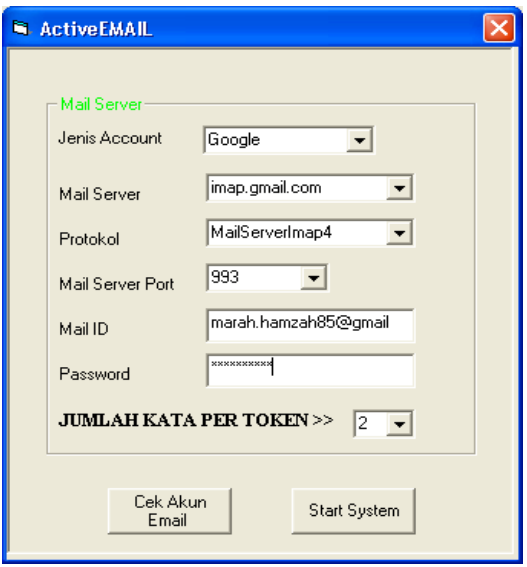
Token dari email dan hasil perhitungannya.

	Token		Peluang_Spam	Peluang_Legitimte
	is outstanding,	16 b...	0	0
	but it	7 b...	0	1
	still has	10 b...	0	0
	serious □□>	12 b...	0	0
	> limitations	14 b...	0	0
	to be	6 b...	0	5
	an Outlook	11 b...	0	0
	replacement and	16 b...	0	0
	we are	7 b...	2	0
	looking to	11 b...	0	0
	□□> >	6 b...	0	14
	replace many	13 b...	0	0
	thousands of	13 b...	0	0
	Outlook installations	22 b...	0	0
	with this	10 b...	1	0

Gambar 4.12 Table Peluang Email

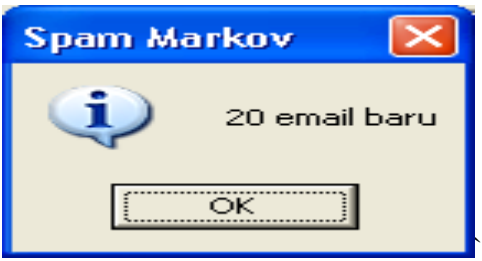
Dari token yang di dapat jumlah probabilitas spam pada email tersebut adalah 19 karena token dari email tersebut terdapat 19 token yang sama dalam tabel spamtoken yang dibentuk pada proses taining, sedangkan dalam tabel ligamentoken terdapat 110 token yang menyerupai maka jumlah probabilitas legitimate adalah 110, untuk penentuan spam kita haus menambah kan nilai π pada probabilitas legitimate maka 112 masih lebih besar dari 19 dan email tersebut adalah legitimate. Setelah email dapat terklasifikasi maka aplikasi dinyatakan telah berhasil dibuat namun, penulis

ingin mengimplementasikan iplikasi ini dengan email yang didapat langsung dari internet. Dimana email didownload langsung dari internet yang sebelumnya telah dikoneksi dengan mengisi Id dan password dari acaount pengguna. Kemudian aplikasi memfilter email yang telah didownload dengan token yang telah dibuat sebelumnya dan system akan memasukkan email Legitimate kedalam Inbox folder yang disediakan oleh system dan kedalam folder Spam jika email tersebut adalah spam.



Gambar 4.13 Active Email

Pengkoneksian account Id secara online

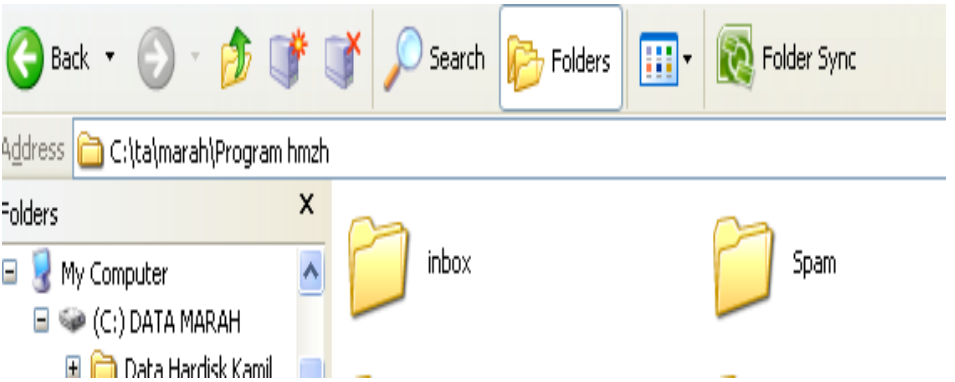


Gambar 4.14 Pemberitahuan Email Baru

Akan ada pemberitahuan jumlah email baru yang masuk pada email tersebut

Gambar 4.15 Pemberitahuan Hasil Klasifikai Email Live

Kemudian system akan memasukkan email tersebut kedalam folder inbox atau spam.



Gambar 4.16 Forder imbox dan spam

Email akan dimasukkan dalam inbox jika dia Legitimate dan spam jika dia spam. Dari hasil penelitian didapat tingkat akurasi 69% karena dari 50 test dataset 13 diantaranya adalah spam sedangkan sistem dapat mendeteksi 9 email jadi 9/13*100% = 69% . dan setelah dilakukan peninjauan lebih lanjut terhadap email test_00010, test_00011, test_00018, test_00041 dimana email tersebut adalah email yang tidak dideteksi oleh sistem sebagai spam diketahui bahwa email tersebut memiliki isi email yang sedikit kata-kata ataupun lebih banyak menggunakan gambar dan hal itulah yang menyebabkan sistem susah mendeteksi email tersebut.

KESIMPULAN

Berikut adalah beberapa kesimpulan yang penulis ambil dari penelitian ini:

1. Metode markov telah terbukti dapat mengklasifikasikan email.
2. Untuk meningkatkan akurasi dari pemfilteran lebih baik menggunakan spam email dari account yang akan difilter sebagai dataset training.
3. Semakin besar ukuran token akan sangat mempengaruhi hasil akurasi pengkalsifikasian email.

REFERENSI

Androutsopoulos, Ion. et al.,1998, *An Experimental Comparison of Naïve Bayesian and Keyword -Based Anti-Spam Filtering with Personal Email Messages*., National Centre for Scientific research Demokritos, Athens., Greece.

Basuki Ahmad, 2006, ***Algoritma Pemograman 2 Menggunakan Visual Basic 6.0***, Institut Teknologi Sepuluh November, Surabaya.

Chandraleka. 2009. "*Cara Mudah Mengelola Email*". MediaKita , Jakarta

E. Walpole, Ronald, 1993, *Pengantar Statistika, Edisi ke-3*, Jakarta: PT Gramedia Pustaka Utama.

Frieyadie, 2010, ***Mudah Belajar Pemograman Database MYSQL dengan Microsft Visual Basic 6.0***, Penerbit Andi, Yogyakarta.

Kadir Abdul, 2008, ***Belajar Database Menggunakan MYSQL***, Penerbit Andi, Yogyakarta.

Kusumadewi.S, 2003, ***Artificial Intelligence (Teknik dan Aplikasinya)***, Penerbit Graha Ilmu, Yoqyakarta

Rusmawan Uus, 2011, ***Visual Basic Untuk semua Tingkatan***, PT. Elex Media Komputindo, Jakarta

Dahliar Ananda, 2011, http://digilib.ittelkm.ac.id/index.php?option= digilib.ittelkom.ac.id/index.php?option=com_repository&Itemid=34 &task=detail&nim=113000093, di unduh tanggal 26 November 213.