

PENGELASAN EMAIL MENGIKUT KATEGORI MENGGUNAKAN
SUPPORT VECTOR MACHINE (SVM)

MARIAH BINTI MOHD DAUD

Laporan projek ini dikemukakan
sebagai memenuhi sebahagian daripada syarat
penganugerahan ijazah Sarjana Muda Sains Komputer

Fakulti Sains Komputer Dan Sistem Maklumat
Universiti Teknologi Malaysia

OKTOBER, 2004

PENGHARGAAN

“ Dengan Nama Allah Yang Maha Pemurah Lagi Maha Mengasihani”

Syukur alhamdulillah, setinggi-tinggi pujian dan kesyukuran saya panjatkan ke hadrat Ilahi kerana dengan izin dan rahmatNya, dapat juga saya menyiapkan projek sarjana muda ini dengan sempurna.

Pertama sekali saya ingin merakamkan jutaan terima kasih kepada penyelia pertama saya iaitu Encik Ahmad Fariz bin Ali dan penyelia kedua saya iaitu Dr Ali bin Selamat yang telah memberikan segala tunjuk ajar dan bimbingan sepanjang proses menyiapkan projek sarjana muda ini. Tidak lupa juga kepada En Norizam Bin Katmon yang memberi kerjasama untuk menyiapkan projek ini. Sesungguhnya segala idea dan dorongan yang diberikan amat berguna dan dihargai.

Jutaan terima kasih buat kedua ibu bapa dan keluarga yang sentiasa memberikan sokongan dan bantuan. Semoga kasih sayang yang terjalin akan bertambah erat. Penghargaan juga ditujukan kepada rakan-rakan seperjuangan yang telah banyak membantu memberikan semangat dan dorongan sehingga berjaya menyiapkan projek ini. Semoga ukhuwah yang terjalin antara kita berkekalan selamanya.

Akhir sekali, penghargaan juga buat semua pihak yang terlibat secara langsung mahupun tidak langsung di dalam menjayakan projek ini. Hanya Allah S.W.T. yang dapat membalas jasa kalian. Wallahua'lam.

ABSTRAK

Kajian tentang bidang pengkategorian teks melibatkan proses pengelasan dokumen teks ke dalam beberapa kategori yang telah ditakrifkan oleh pengguna. Objektif bagi projek ini ialah untuk membuat kajian proses pengelasan email mengikut kategori dengan menggunakan perisian Support Vector Machine (SVM). Antara proses yang digunakan ialah membaca data input email dari bahagian subjek dan *body*, mengekstrakan ciri, pemilihan ciri dan mengelaskan data menggunakan SVM. Proses mengekstrakan ciri melibatkan kaedah *word stopping* dan *word stemming* di mana dapat mengurangkan bilangan dimensi ciri. Proses pemilihan ciri melibatkan kaedah *TFIDF*. Keberkesanan pengelasan diukur menggunakan kriteria *precision* dan *recall*. Keputusan yang terhasil daripada analisis menunjukkan di mana Support Vector Machine sangat efektif dalam proses pengelasan email.

ABSTRACT

Study on text categorization field contains classification process of text documents into a fixed number of pre-defined categories by user. The objective of this project is to make research on classifying email process based on category using Support Vector Machine software. Among processes will be used are read input data email from subject and body, feature extraction, feature selection and classify data using Support Vector Machine (SVM). Feature extraction process involved word stopping and word stemming methods that can reduce the number of dimension of features. Features selection process involved *TFIDF* method. Effectiveness of classification process has been measured using precision and recall criteria. Result produced from analysis showed that Support Vector Machine is very effective in email classifying process.

KANDUNGAN

BAB	PERKARA	HALAMAN
	JUDUL	i
	PENGAKUAN	ii
	DEDIKASI	iii
	PENGHARGAAN	iv
	ABSTRAK	v
	ABSTRACT	vi
	KANDUNGAN	vii
	SENARAI RAJAH	xiii
	SENARAI JADUAL	xv
	SENARAI LAMPIRAN	xvi
	SENARAI ISTILAH	xvii
	SENARAI SIMBOL	xviii
	SENARAI RUMUS	xix
	SENARAI SINGKATAN	xx
1	Pengenalan	
	1.1 Pendahuluan	1
	1.2 Penyataan Masalah	3
	1.3 Matlamat	4

1.4	Objektif	4
1.5	Skop	5
1.6	Penyelesaian Masalah	6
1.7	Kekangan dan Had Limitasi	7
1.8	Justifikasi dan Kepentingan	7
1.9	Kesimpulan	8
2	KAJIAN LITERATUR	
2.1	Pengenalan	9
2.2	Pengelasan	10
	2.2.1 Pengelasan secara manual	10
	2.2.2 Pengelasan secara automatik	10
2.3	Pengenalan Kategori Teks	11
	2.3.1 Manual	12
	2.3.2 Berdasarkan Aturan (<i>Rule-Based</i>)	13
	2.3.2 Pembelajaran Melalui Seliaan (<i>Supervised Learnig</i>)	14
	2.3.4 Pembelajaran Tanpa Seliaan (<i>Unsupervised Learning</i>)	14
	2.3.5 Perbandingan Pendekatan	16
2.4	Penapisan Email	17
2.5	Konsep Ciri (<i>Feature</i>)	19
2.6	Kaedah <i>TFIDF</i>	22
2.7	Pengelasan Teks Email	23
2.8	Kajian Ke atas Teknik Support Vector Machine(SVM)	24
2.9	Teknik SVM	25
	2.9.1 Pengelas Linear	26
	2.9.2 Pengelas Tidak Linear (non-linear)	29
	2.9.3 Jenis Kernel Pengelas Linear	32
2.10	Ciri-ciri SVM	33

2.11	Kelebihan dan kekurangan SVM	35
2.12	Kajian Perisian SVM	36
2.13	Teknik Lain Bagi Pengelasan Email	37
	2.13.1 Rangkaian Neural (Neural Network)	37
	2.13.2 k -Nearest Neighbor (k -NN)	39
2.14	Perbandingan Teknik-teknik Pengelasan	41
2.15	Kajian Ke Atas Projek Pengelasan Email Terdahulu	43
2.16	Kesimpulan	44

3 METODOLOGI KAJIAN

3.1	Pengenalan	46
3.2	Analisa Keperluan	46
	3.2.1 Keperluan Perisian	47
	3.2.2 Bahasa Pengaturcaraan	47
	3.2.3 Keperluan Perkakasan Minima	47
	3.2.4 Justifikasi Perkakasan	48
3.3	Proses Pengelasan Email	49
3.4	Algoritma Pengelasan Data Email	51
3.5	Pemprosesan Dokumen Email	53
3.6	Pengekstrakan Ciri (Feature Extraction)	53
3.7	Pemilihan Ciri (Feature Selection)	56
3.8	Perwakilan Kandungan Email (ciri vector)	56
3.9	Proses Pengumpulan Data	57
3.10	Format Input Dalam Bentuk <i>tfidf</i>	57
3.11	Format Fail Melatih dan Menguji	58
3.12	Proses Melatih SVM	59
3.13	Proses Menguji SVM	60
3.14	Kriteria Penilaian Keberkesanan (<i>Relevance Evaluation</i>)	60

3.15	Kekangan dan Limitasi	62
3.16	Kesimpulan	62
4	IMPLEMENTASI	
4.1	Pengenalan	63
4.2	Aliran Kerja Proses Pengelasan	64
	4.2.1 Penerangan Aliran Proses Pengelasan	65
4.3	Proses Capaian Email	65
4.4	Proses Membaca Input	67
4.5	Proses Token Perkataan	68
4.6	Proses <i>Stopping</i>	69
4.7	Proses <i>Stemming</i>	70
4.8	Proses Mendapatkan Nilai <i>tfidf</i> Sebagai Input SVM	73
4.9	Format Input Bentuk <i>tfidf</i>	76
4.10	Pengelasan Menggunakan Perisian SVM	77
4.11	Penilaian Keberkesanan	78
4.12	Kesimpulan	79
5	HASIL KAJIAN, PENGUJIAN DAN PENCAPAIAN	
5.1	Pengenalan	80
5.2	Pengelasan Email	80
5.3	Input Mesej Email Dalam Platform Microsoft Outlook	81
5.4	Proses Pertukaran Format Mesej Email	82
5.5	Input Dan Output Selepas Proses Token	83
5.6	Proses <i>Stopping</i>	84

5.7	Proses <i>Stemming</i>	85
5.8	Proses Mencari <i>tfidf</i> Sebagai Input SVM	86
	5.8.1 <i>Term Frequency, tf</i>	86
	5.8.2 <i>Document Frequency, df</i>	88
	5.8.3 <i>Inverse Document Frequency, idf</i>	88
	5.8.4 <i>Term Frequency * Inverse Document Frequency</i>	88
5.9	Pengelasan Menggunakan Perisian SVM	91
5.10	Penilaian Keberkesanan (<i>Precision</i> dan <i>Recall</i>)	94
5.11	Hasil Pengujian Pengelasan Menggunakan Perisian SVM	94
	5.11.1 <i>Precision</i> (Keberkesanan)	95
	5.11.2 <i>Recall</i>	97
	5.11.3 <i>Accuracy</i> (Ketepatan)	98
	5.11.4 Kadar Ralat (<i>Error Rate</i>)	100
	5.11.5 Purata Nilai <i>Precision</i> dan <i>Recall</i> Setiap Kategori	101
	5.11.5.1 Kategori Design	102
	5.11.5.2 Kategori Friend	103
	5.11.5.3 Kategori Love	104
	5.11.5.4 Kategori Web	105
	5.11.5.5 Kategori Wanita lelaki	106
	5.11.5.6 Kategori Cogramm	107
	5.11.5.7 Kategori Job	108
5.12	Perbincangan Keberkesanan Perisian Yang Dipilih	109
5.13	Kesimpulan	110

6 PERBINCANGAN DAN KESIMPULAN

6.1	Pengenalan	111
6.2	Pencapaian	111
6.3	Analisa Input	112

6.4	Analisa Pengekstrakan Ciri (<i>Stopping dan Stemming</i>)	113
6.5	Analisa Perisian Yang Digunakan	114
6.6	Masalah Yang Dihadapi	114
6.7	Cadangan Pembaikan	115
6.8	Kesimpulan	116

RUJUKAN	117
Lampiran A - F	123 - 141

BAB 1

PENGENALAN

1.1 Pendahuluan

Pada masa kini, internet dan komputer sangat penting dalam kehidupan. Pelbagai cara digunakan untuk berkomunikasi dengan orang perseorangan. Pada masa dahulu, setiap orang berhubung antara satu sama lain dengan hanya menghantar surat melalui pos. Proses ini memakan masa yang agak lama untuk menghantar sesuatu mesej kepada rakan kita. Setelah telefon diperkenalkan, ramai orang menggunakan telefon untuk berhubung antara satu sama lain.

Dengan ini, pembangun laman web memikirkan cara bagaimana untuk menyelesaikan masalah yang berlaku sekiranya ingin berhubung dengan orang yang jauh. Satu pendekatan lain digunakan iaitu menggunakan elektronik mail (e-mail) untuk berhubung antara satu sama lain. Pendekatan ini sangat mudah, menjimatkan kos dan masa. Walaupun pendekatan baru ini digunakan pada masa kini, perkhimatan surat melalui pos masih mempunyai kepentingannya [31].

Aplikasi e-mail telah berkembang dengan begitu pesat sekali. Bilangan pengguna yang ramai telah menjadikan aplikasi ini semakin maju. Setiap pengguna laman web boleh mempunyai akaun email sendiri. Aplikasi email memberi banyak faedah kepada pengguna untuk memudahkan perhubungan dan komunikasi antara setiap pengguna melalui web. Sekiranya pengguna ingin mempunyai akaun email, pengguna mesti mendaftar terlebih dahulu di mana-mana laman web yang menyediakan perkhidmatan email contohnya *yahoo*, *hotmail*, *lycos* dan banyak lagi.

Akaun pengguna (penerima email) akan mengandungi semua maklumat email penghantar yang menghantar email kepada penerima tersebut. Kesemua maklumat email yang diterima akan dimasukkan ke dalam inbox iaitu satu *folder* yang menyediakan fungsi untuk menyimpan semua maklumat email yang sampai ke dalam akaun email pengguna.

Sekiranya terlalu banyak email yang diterima telah diletakkan dalam inbox, pengguna keliru untuk membaca email mana terlebih dahulu. Pengguna tidak sempat untuk mengemaskini akaun sekiranya terlalu banyak menerima email dalam sehari. Maka, pendekatan pengelasan kategori email dikaji untuk memudahkan pengguna membaca email dengan mudah mengikut kategori yang telah ditetapkan. Pengguna akan *create folder* secara manual dan maklumat email yang berkaitan dengan tajuk *folder* tersebut akan terus dimasukkan dalam *folder* tersebut secara automatik. Sekiranya maklumat email yang sampai tidak tersenarai dalam kategori yang ada, maklumat email tersebut akan dikategorikan dalam inbox sahaja.

Pendekatan pengelasan kategori email ini dapat membantu pengguna email untuk membaca email dengan lebih mudah. Selalunya, pengguna akan mengelaskan akaun email mengikut nama penghantar email, tajuk email dan sebagainya. Contoh bagi

atribut pensyarah, email yang sering dikelaskan oleh pensyarah ialah mengikut kategori iaitu ‘pelajar’, ‘pejabat am, subjek, pensyarah lain dan sebagainya.

1.2 Penyataan Masalah

Dalam sehari, pengguna email menerima begitu banyak email. Semua email tersebut akan dimasukkan terus ke dalam inbox pengguna email. Jadi, ini akan menjadikan kandungan dalam inbox menjadi tidak tersusun dan pengguna berasa keliru untuk melihat manakah email yang perlu dibaca terlebih dahulu. Pengguna juga susah untuk mengenal pasti maklumat email tersebut berada dalam kategori mana.

Masalah lain dalam email ialah *too much* bermaksud terlalu banyak email yang diterima dalam inbox. Jadi, bagaimana kita hendak meluangkan masa untuk membaca email yang begitu banyak dengan pantas. Mungkin kita perlu meluangkan lebih banyak masa untuk membaca email tersebut. Sejak meningkatnya kefungisian berkomunikasi melalui email, masa untuk kita menjawab email juga perlu diambil kira [17].

Bagi masalah email yang dihadapi sekarang, penyelesaiannya ialah pengelasan email mengikut kategori. Selama ini, email yang diterima setiap hari hanya dimasukkan ke dalam inbox. Sekiranya pengguna telah sedia ada *folder* mengikut kategori tertentu, maklumat email yang sampai tetap dihantar ke dalam *inbox* tanpa mengikut kategori. Jadi, sekiranya pengguna hendak mengemaskinikan email mengikut kategori, pengguna akan mengemaskinikan maklumat email tersebut secara manual.

1.3 Matlamat

Matlamat utama projek ini ialah untuk membuat kajian proses pengelasan email menggunakan perisian Support Vector Machine (SVM) dan seterusnya menganalisa keberkesanan perisian tersebut.

1.4 Objektif

Kajian yang dilakukan ini adalah untuk memenuhi objektif-objektif berikut:-

- i) Membuat kajian proses pengelasan email mengikut kategori menggunakan perisian Support Vector Machine
- ii) Mengumpul data email sebenar untuk melaksanakan proses latihan (*training*) bagi menguji keberkesanan perisian Support Vector Machine(SVM) dalam proses menyelesaikan masalah pengelasan
- iii) Proses mendapatkan nilai *tfidf* dilakukan untuk dijadikan input bagi perisian Support Vector Machine
- iv) Keberkesanan teknik pengelasan email menggunakan Support Vector Machine (SVM) dinilai berdasarkan pendekatan *precision* dan *recall*

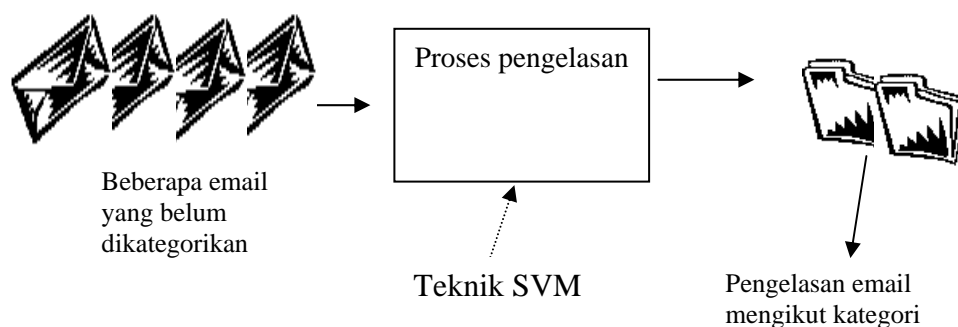
1.5 Skop Projek

Antara skop projek ialah :-

- i) Perisian ini akan melaksanakan tugas menerima input (email), menganalisa dan mengelaskan output (email) kepada kategori email yang telah ditetapkan
- ii) Pengelasan email dibuat berdasarkan kepada bahagian *subject* dan *body* yang terdapat dalam kandungan email
- iii) Pengelasan email dilakukan menggunakan perisian Support Vector Machine (SVM) dari SVM^{light}
- iv) Aplikasi pengelasan akan menggunakan Microsoft Outlook sebagai platform untuk proses input data email
- v) Kelas atau kategori pengelasan yang dilakukan ialah kategori cogramm, wanita lelaki, job, web, design, friend dan love
- vi) Jumlah email yang digunakan semasa proses *training* (pembelajaran) adalah sebanyak 200 data email dan semasa proses *testing* (pengujian), sebanyak 100 data email
- vii) Proses pengelasan hanya dilakukan ke atas email yang mengandungi teks sahaja
- viii) Proses *stemming* melibatkan bahasa melayu dan bahasa inggeris

1.6 Penyelesaian Masalah

Projek ini dijalankan bertujuan untuk menyelesaikan masalah pengelasan kategori email. Kandungan mesej baru yang terlalu banyak dalam senarai email amat menyusahkan pengguna. Pengguna keliru untuk menilai manakah email yang lebih berkepentingan dan perlu dibaca terlebih dahulu. Projek ini dijalankan untuk melihat sejauh mana teknik pengelasan yang dipilih dapat membantu menguruskan email dengan lebih efisien iaitu melalui proses pengelasan email.



Rajah 1.1 : Gambaran kasar pengelasan kategori email

Contoh gambaran penyelesaian dapat dilihat seperti di atas tentang bagaimana penyelesaian masalah dapat dibuat. Pertama sekali email yang ada dalam *folder* inbox adalah email yang belum dikategorikan. Kemudian, email-email tersebut akan melalui proses pengelasan supaya email-email tersebut akan dikategorikan mengikut setiap kategori yang telah ditetapkan. Contoh kategori email yang dibuat oleh pensyarah ialah subjek, pejabat am, pelajar dan sebagainya.

1.7 Kekangan dan Had Limitasi

Antara kekangan dan had limitasi semasa proses kajian dilakukan ialah :-

- i. Bilangan input data yang akan diuji agak terbatas.
- ii. Perisian SVM^{light} hanya digunakan untuk pengujian proses pengelasan.
- iii. Kategori email ditakrif terlebih dahulu.
- iv. Setiap input email yang digunakan perlu ditakrif terlebih dahulu dalam program yang digunakan untuk mendapatkan nilai *tfidf*
- v. Ketepatan untuk mengkategorikan email kemungkinan tidak mencapai 100%

1.8 Justifikasi dan Kepentingan

Kelebihan setelah melakukan kajian proses pengelasan email :-

- i) Dapat mengelaskan email mengikut kategori yang ditetapkan.
- ii) Projek ini dijalankan untuk mengkaji bagaimana Support Vector Machine (SVM) berdasarkan perisian SVM^{light} boleh digunakan untuk menyelesaikan masalah pengelasan email.
- iii) Pengelasan dilakukan ke atas kandungan email di dalam bahasa melayu dan bahasa inggeris

1.9 Kesimpulan

Pada masa kini, ramai pengguna menggunakan email untuk berkomunikasi dan menghantar mesej antara satu sama lain. Proses pengelasan email memudahkan pengguna membaca maklumat email yang sampai ke dalam akaun email. Kajian yang dilakukan ini akan membantu pembangun lain untuk membangunkan satu prototaip yang dapat melakukan kerja pengelasan email mengikut kategori yang ditetapkan. Adalah diharapkan kajian pengelasan email ini dapat dijadikan permulaan kepada penghasilan pakej yang lebih mantap dan meluas.

PENGELASAN EMAIL MENGIKUT KATEGORI MENGGUNAKAN
SUPPORT VECTOR MACHINE (SVM)

MARIAH BINTI MOHD DAUD

Laporan projek ini dikemukakan
sebagai memenuhi sebahagian daripada syarat
penganugerahan ijazah Sarjana Muda Sains Komputer

Fakulti Sains Komputer Dan Sistem Maklumat
Universiti Teknologi Malaysia

OKTOBER, 2004

UNIVERSITI TEKNOLOGI MALAYSIA

BORANG PENCESAHAN STATUS TESIS ♦

JUDUL : PENGELASAN EMAIL MENGIKUT KATEGORI MENGGUNAKAN
SUPPORT VECTOR MACHINE (SVM)

SESI PENGAJIAN : SEMESTER I 2004/2005

Saya,

MARIAH BINTI MOHD DAUD

(HURUF BESAR)

mengaku membenarkan tesis (PSM/Sarjana/Doktor Falsafah)* ini disimpan di Perpustakaan Universiti Teknologi Malaysia dengan syarat-syarat kegunaan seperti berikut :

1. Tesis adalah hakmilik Universiti Teknologi Malaysia.
2. Perpustakaan Universiti teknologi Malaysia dibenarkan membuat salinan untuk tujuan pengajian sahaja.
3. Perpustakaan dibenarkan membuat salinan tesis ini sebagai bahan pertukaran antara institusi pengajian tinggi.
4. **Sila Tandakan (√)

SULIT

(Mengandungi maklumat yang berdarjah keselamatan atau kepentingan Malaysiayang termaktub didalam AKTA RAHSIA RASMI 1972)

TERHAD

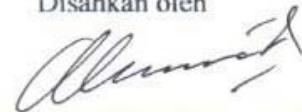
(Mengandungi maklumat TERHAD yang telah ditentukan oleh organisasi/badan dimana penyelidikan dijalankan)

TIDAK TERHAD

Disahkan oleh



(TANDATANGAN PENULIS)



(TANDATANGAN PENYELIA)

Alamat tetap:

B 33 KG MATANG GELOK,
PEKAN AYER HITAM,
06150, ALOR STAR, KEDAH

ENCIK AHMAD FARIZ BIN ALI

Nama Penyelia

Tarikh : 18 OKTOBER 2004

Tarikh: 18 OKTOBER 2004


- CATATAN
- * Potong yang tidak berkenaan
 - ** Jika tesis ini SULIT atau TERHAD, sila lampirkan surat daripada pihak berkuasa/organisasi berkenaan dengan menyatakan sekali sebab dan tempoh tesis ini perlu dikelaskan sebagai SULIT atau TERHAD
 - ♦ Tesis dimaksudkan sebagai tesis bagi Ijazah Doktor Falsafah dan Sarjana secara penyelidikan, atau disertasi bagi pengajian kerja kursus dan penyelidikan, atau Laporan Projek Sarjana Muda


“Saya akui karya ini adalah hasil kerja saya sendiri kecuali nukilan dan ringkasan yang tiap-tiap satunya telah saya jelaskan sumbernya”.



Tandatangan :
Nama Penulis : MARIAH BINTI MOHD DAUD
Tarikh : 18 OKTOBER 2004

“Kami akui bahawa kami telah membaca karya ini dan pada pandangan kami karya ini adalah memadai dari segi skop dan kualiti untuk tujuan penganugerahan ijazah Sarjana Muda Sains Komputer”.

Tandatangan : 
Nama Penyelia I : ENCIK AHMAD FARIZ BIN ALI
Tarikh : 18 OKTOBER 2004

Tandatangan : 
Nama Penyelia II : DR. ALI BIN SELAMAT
Tarikh : 18 OKTOBER 2004

PENGHARGAAN

“ Dengan Nama Allah Yang Maha Pemurah Lagi Maha Mengasihani”

Syukur alhamdulillah, setinggi-tinggi pujian dan kesyukuran saya panjatkan ke hadrat Ilahi kerana dengan izin dan rahmatNya, dapat juga saya menyiapkan projek sarjana muda ini dengan sempurna.

Pertama sekali saya ingin merakamkan jutaan terima kasih kepada penyelia pertama saya iaitu Encik Ahmad Fariz bin Ali dan penyelia kedua saya iaitu Dr Ali bin Selamat yang telah memberikan segala tunjuk ajar dan bimbingan sepanjang proses menyiapkan projek sarjana muda ini. Tidak lupa juga kepada En Norizam Bin Katmon yang memberi kerjasama untuk menyiapkan projek ini. Sesungguhnya segala idea dan dorongan yang diberikan amat berguna dan dihargai.

Jutaan terima kasih buat kedua ibu bapa dan keluarga yang sentiasa memberikan sokongan dan bantuan. Semoga kasih sayang yang terjalin akan bertambah erat. Penghargaan juga ditujukan kepada rakan-rakan seperjuangan yang telah banyak membantu memberikan semangat dan dorongan sehingga berjaya menyiapkan projek ini. Semoga ukhuwah yang terjalin antara kita berkekalan selamanya.

Akhir sekali, penghargaan juga buat semua pihak yang terlibat secara langsung mahupun tidak langsung di dalam menjayakan projek ini. Hanya Allah S.W.T. yang dapat membalas jasa kalian. Wallahua'lam.

ABSTRAK

Kajian tentang bidang pengkategorian teks melibatkan proses pengelasan dokumen teks ke dalam beberapa kategori yang telah ditakrifkan oleh pengguna. Objektif bagi projek ini ialah untuk membuat kajian proses pengelasan email mengikut kategori dengan menggunakan perisian Support Vector Machine (SVM). Antara proses yang digunakan ialah membaca data input email dari bahagian subjek dan *body*, pengekstrakan ciri, pemilihan ciri dan mengelaskan data menggunakan SVM. Proses pengekstrakan ciri melibatkan kaedah *word stopping* dan *word stemming* di mana dapat mengurangkan bilangan dimensi ciri. Proses pemilihan ciri melibatkan kaedah *TFIDF*. Keberkesanan pengelasan diukur menggunakan kriteria *precision* dan *recall*. Keputusan yang terhasil daripada analisis menunjukkan di mana Support Vector Machine sangat efektif dalam proses pengelasan email.

ABSTRACT

Study on text categorization field contains classification process of text documents into a fixed number of pre-defined categories by user. The objective of this project is to make research on classifying email process based on category using Support Vector Machine software. Among processes will be used are read input data email from subject and body, feature extraction, feature selection and classify data using Support Vector Machine (SVM). Feature extraction process involved word stopping and word stemming methods that can reduce the number of dimension of features. Features selection process involved *TFIDF* method. Effectiveness of classification process has been measured using precision and recall criteria. Result produced from analysis showed that Support Vector Machine is very effective in email classifying process.

KANDUNGAN

BAB	PERKARA	HALAMAN
	JUDUL	i
	PENGAKUAN	ii
	DEDIKASI	iii
	PENGHARGAAN	iv
	ABSTRAK	v
	ABSTRACT	vi
	KANDUNGAN	vii
	SENARAI RAJAH	xiii
	SENARAI JADUAL	xv
	SENARAI LAMPIRAN	xvi
	SENARAI ISTILAH	xvii
	SENARAI SIMBOL	xviii
	SENARAI RUMUS	xix
	SENARAI SINGKATAN	xx
1	Pengenalan	
	1.1 Pendahuluan	1
	1.2 Penyataan Masalah	3
	1.3 Matlamat	4

1.4	Objektif	4
1.5	Skop	5
1.6	Penyelesaian Masalah	6
1.7	Kekangan dan Had Limitasi	7
1.8	Justifikasi dan Kepentingan	7
1.9	Kesimpulan	8
2	KAJIAN LITERATUR	
2.1	Pengenalan	9
2.2	Pengelasan	10
	2.2.1 Pengelasan secara manual	10
	2.2.2 Pengelasan secara automatik	10
2.3	Pengenalan Kategori Teks	11
	2.3.1 Manual	12
	2.3.2 Berdasarkan Aturan (<i>Rule-Based</i>)	13
	2.3.2 Pembelajaran Melalui Seliaan (<i>Supervised Learnig</i>)	14
	2.3.4 Pembelajaran Tanpa Seliaan (<i>Unsupervised Learning</i>)	14
	2.3.5 Perbandingan Pendekatan	16
2.4	Penapisan Email	17
2.5	Konsep Ciri (<i>Feature</i>)	19
2.6	Kaedah <i>TFIDF</i>	22
2.7	Pengelasan Teks Email	23
2.8	Kajian Ke atas Teknik Support Vector Machine(SVM)	24
2.9	Teknik SVM	25
	2.9.1 Pengelas Linear	26
	2.9.2 Pengelas Tidak Linear (non-linear)	29
	2.9.3 Jenis Kernel Pengelas Linear	32
2.10	Ciri-ciri SVM	33

2.11	Kelebihan dan kekurangan SVM	35
2.12	Kajian Perisian SVM	36
2.13	Teknik Lain Bagi Pengelasan Email	37
	2.13.1 Rangkaian Neural (Neural Network)	37
	2.13.2 k -Nearest Neighbor (k -NN)	39
2.14	Perbandingan Teknik-teknik Pengelasan	41
2.15	Kajian Ke Atas Projek Pengelasan Email Terdahulu	43
2.16	Kesimpulan	44

3 METODOLOGI KAJIAN

3.1	Pengenalan	46
3.2	Analisa Keperluan	46
	3.2.1 Keperluan Perisian	47
	3.2.2 Bahasa Pengaturcaraan	47
	3.2.3 Keperluan Perkakasan Minima	47
	3.2.4 Justifikasi Perkakasan	48
3.3	Proses Pengelasan Email	49
3.4	Algoritma Pengelasan Data Email	51
3.5	Pemprosesan Dokumen Email	53
3.6	Pengekstrakan Ciri (Feature Extraction)	53
3.7	Pemilihan Ciri (Feature Selection)	56
3.8	Perwakilan Kandungan Email (ciri vector)	56
3.9	Proses Pengumpulan Data	57
3.10	Format Input Dalam Bentuk <i>tfidf</i>	57
3.11	Format Fail Melatih dan Menguji	58
3.12	Proses Melatih SVM	59
3.13	Proses Menguji SVM	60
3.14	Kriteria Penilaian Keberkesanan (<i>Relevance Evaluation</i>)	60

3.15	Kekangan dan Limitasi	62
3.16	Kesimpulan	62
4	IMPLEMENTASI	
4.1	Pengenalan	63
4.2	Aliran Kerja Proses Pengelasan	64
	4.2.1 Penerangan Aliran Proses Pengelasan	65
4.3	Proses Capaian Email	65
4.4	Proses Membaca Input	67
4.5	Proses Token Perkataan	68
4.6	Proses <i>Stopping</i>	69
4.7	Proses <i>Stemming</i>	70
4.8	Proses Mendapatkan Nilai <i>tfidf</i> Sebagai Input SVM	73
4.9	Format Input Bentuk <i>tfidf</i>	76
4.10	Pengelasan Menggunakan Perisian SVM	77
4.11	Penilaian Keberkesanan	78
4.12	Kesimpulan	79
5	HASIL KAJIAN, PENGUJIAN DAN PENCAPAIAN	
5.1	Pengenalan	80
5.2	Pengelasan Email	80
5.3	Input Mesej Email Dalam Platform Microsoft Outlook	81
5.4	Proses Pertukaran Format Mesej Email	82
5.5	Input Dan Output Selepas Proses Token	83
5.6	Proses <i>Stopping</i>	84

5.7	Proses <i>Stemming</i>	85
5.8	Proses Mencari <i>tfidf</i> Sebagai Input SVM	86
	5.8.1 <i>Term Frequency, tf</i>	86
	5.8.2 <i>Document Frequency, df</i>	88
	5.8.3 <i>Inverse Document Frequency, idf</i>	88
	5.8.4 <i>Term Frequency * Inverse Document Frequency</i>	88
5.9	Pengelasan Menggunakan Perisian SVM	91
5.10	Penilaian Keberkesanan (<i>Precision</i> dan <i>Recall</i>)	94
5.11	Hasil Pengujian Pengelasan Menggunakan Perisian SVM	94
	5.11.1 <i>Precision</i> (Keberkesanan)	95
	5.11.2 <i>Recall</i>	97
	5.11.3 <i>Accuracy</i> (Ketepatan)	98
	5.11.4 Kadar Ralat (<i>Error Rate</i>)	100
	5.11.5 Purata Nilai <i>Precision</i> dan <i>Recall</i> Setiap Kategori	101
	5.11.5.1 Kategori Design	102
	5.11.5.2 Kategori Friend	103
	5.11.5.3 Kategori Love	104
	5.11.5.4 Kategori Web	105
	5.11.5.5 Kategori Wanita lelaki	106
	5.11.5.6 Kategori Cogramm	107
	5.11.5.7 Kategori Job	108
5.12	Perbincangan Keberkesanan Perisian Yang Dipilih	109
5.13	Kesimpulan	110

6 PERBINCANGAN DAN KESIMPULAN

6.1	Pengenalan	111
6.2	Pencapaian	111
6.3	Analisa Input	112

6.4	Analisa Pengekstrakan Ciri (<i>Stopping dan Stemming</i>)	113
6.5	Analisa Perisian Yang Digunakan	114
6.6	Masalah Yang Dihadapi	114
6.7	Cadangan Pembaikan	115
6.8	Kesimpulan	116

RUJUKAN	117
Lampiran A - F	123 - 141

BAB 1

PENGENALAN

1.1 Pendahuluan

Pada masa kini, internet dan komputer sangat penting dalam kehidupan. Pelbagai cara digunakan untuk berkomunikasi dengan orang perseorangan. Pada masa dahulu, setiap orang berhubung antara satu sama lain dengan hanya menghantar surat melalui pos. Proses ini memakan masa yang agak lama untuk menghantar sesuatu mesej kepada rakan kita. Setelah telefon diperkenalkan, ramai orang menggunakan telefon untuk berhubung antara satu sama lain.

Dengan ini, pembangun laman web memikirkan cara bagaimana untuk menyelesaikan masalah yang berlaku sekiranya ingin berhubung dengan orang yang jauh. Satu pendekatan lain digunakan iaitu menggunakan elektronik mail (e-mail) untuk berhubung antara satu sama lain. Pendekatan ini sangat mudah, menjimatkan kos dan masa. Walaupun pendekatan baru ini digunakan pada masa kini, perkhimatan surat melalui pos masih mempunyai kepentingannya [31].

Aplikasi e-mail telah berkembang dengan begitu pesat sekali. Bilangan pengguna yang ramai telah menjadikan aplikasi ini semakin maju. Setiap pengguna laman web boleh mempunyai akaun email sendiri. Aplikasi email memberi banyak faedah kepada pengguna untuk memudahkan perhubungan dan komunikasi antara setiap pengguna melalui web. Sekiranya pengguna ingin mempunyai akaun email, pengguna mesti mendaftar terlebih dahulu di mana-mana laman web yang menyediakan perkhidmatan email contohnya *yahoo*, *hotmail*, *lycos* dan banyak lagi.

Akaun pengguna (penerima email) akan mengandungi semua maklumat email penghantar yang menghantar email kepada penerima tersebut. Kesemua maklumat email yang diterima akan dimasukkan ke dalam inbox iaitu satu *folder* yang menyediakan fungsi untuk menyimpan semua maklumat email yang sampai ke dalam akaun email pengguna.

Sekiranya terlalu banyak email yang diterima telah diletakkan dalam inbox, pengguna keliru untuk membaca email mana terlebih dahulu. Pengguna tidak sempat untuk mengemaskini akaun sekiranya terlalu banyak menerima email dalam sehari. Maka, pendekatan pengelasan kategori email dikaji untuk memudahkan pengguna membaca email dengan mudah mengikut kategori yang telah ditetapkan. Pengguna akan *create folder* secara manual dan maklumat email yang berkaitan dengan tajuk *folder* tersebut akan terus dimasukkan dalam *folder* tersebut secara automatik. Sekiranya maklumat email yang sampai tidak tersenarai dalam kategori yang ada, maklumat email tersebut akan dikategorikan dalam inbox sahaja.

Pendekatan pengelasan kategori email ini dapat membantu pengguna email untuk membaca email dengan lebih mudah. Selalunya, pengguna akan mengelaskan akaun email mengikut nama penghantar email, tajuk email dan sebagainya. Contoh bagi

atribut pensyarah, email yang sering dikelaskan oleh pensyarah ialah mengikut kategori iaitu ‘pelajar’, ‘pejabat am, subjek, pensyarah lain dan sebagainya.

1.2 Penyataan Masalah

Dalam sehari, pengguna email menerima begitu banyak email. Semua email tersebut akan dimasukkan terus ke dalam inbox pengguna email. Jadi, ini akan menjadikan kandungan dalam inbox menjadi tidak tersusun dan pengguna berasa keliru untuk melihat manakah email yang perlu dibaca terlebih dahulu. Pengguna juga susah untuk mengenal pasti maklumat email tersebut berada dalam kategori mana.

Masalah lain dalam email ialah *too much* bermaksud terlalu banyak email yang diterima dalam inbox. Jadi, bagaimana kita hendak meluangkan masa untuk membaca email yang begitu banyak dengan pantas. Mungkin kita perlu meluangkan lebih banyak masa untuk membaca email tersebut. Sejak meningkatnya kefungisian berkomunikasi melalui email, masa untuk kita menjawab email juga perlu diambil kira [17].

Bagi masalah email yang dihadapi sekarang, penyelesaiannya ialah pengelasan email mengikut kategori. Selama ini, email yang diterima setiap hari hanya dimasukkan ke dalam inbox. Sekiranya pengguna telah sedia ada *folder* mengikut kategori tertentu, maklumat email yang sampai tetap dihantar ke dalam *inbox* tanpa mengikut kategori. Jadi, sekiranya pengguna hendak mengemaskinikan email mengikut kategori, pengguna akan mengemaskinikan maklumat email tersebut secara manual.

1.3 Matlamat

Matlamat utama projek ini ialah untuk membuat kajian proses pengelasan email menggunakan perisian Support Vector Machine (SVM) dan seterusnya menganalisa keberkesanan perisian tersebut.

1.4 Objektif

Kajian yang dilakukan ini adalah untuk memenuhi objektif-objektif berikut:-

- i) Membuat kajian proses pengelasan email mengikut kategori menggunakan perisian Support Vector Machine
- ii) Mengumpul data email sebenar untuk melaksanakan proses latihan (*training*) bagi menguji keberkesanan perisian Support Vector Machine(SVM) dalam proses menyelesaikan masalah pengelasan
- iii) Proses mendapatkan nilai *tfidf* dilakukan untuk dijadikan input bagi perisian Support Vector Machine
- iv) Keberkesanan teknik pengelasan email menggunakan Support Vector Machine (SVM) dinilai berdasarkan pendekatan *precision* dan *recall*

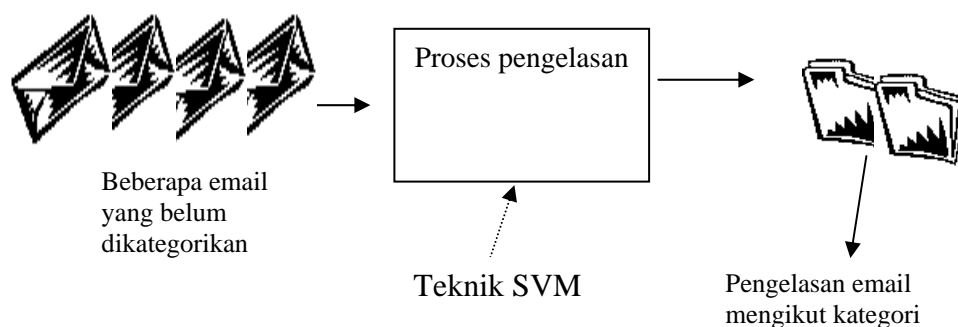
1.5 Skop Projek

Antara skop projek ialah :-

- i) Perisian ini akan melaksanakan tugas menerima input (email), menganalisa dan mengelaskan output (email) kepada kategori email yang telah ditetapkan
- ii) Pengelasan email dibuat berdasarkan kepada bahagian *subject* dan *body* yang terdapat dalam kandungan email
- iii) Pengelasan email dilakukan menggunakan perisian Support Vector Machine (SVM) dari SVM^{light}
- iv) Aplikasi pengelasan akan menggunakan Microsoft Outlook sebagai platform untuk proses input data email
- v) Kelas atau kategori pengelasan yang dilakukan ialah kategori cogramm, wanita lelaki, job, web, design, friend dan love
- vi) Jumlah email yang digunakan semasa proses *training* (pembelajaran) adalah sebanyak 200 data email dan semasa proses *testing* (pengujian), sebanyak 100 data email
- vii) Proses pengelasan hanya dilakukan ke atas email yang mengandungi teks sahaja
- viii) Proses *stemming* melibatkan bahasa melayu dan bahasa inggeris

1.6 Penyelesaian Masalah

Projek ini dijalankan bertujuan untuk menyelesaikan masalah pengelasan kategori email. Kandungan mesej baru yang terlalu banyak dalam senarai email amat menyusahkan pengguna. Pengguna keliru untuk menilai manakah email yang lebih berkepentingan dan perlu dibaca terlebih dahulu. Projek ini dijalankan untuk melihat sejauh mana teknik pengelasan yang dipilih dapat membantu menguruskan email dengan lebih efisien iaitu melalui proses pengelasan email.



Rajah 1.1 : Gambaran kasar pengelasan kategori email

Contoh gambaran penyelesaian dapat dilihat seperti di atas tentang bagaimana penyelesaian masalah dapat dibuat. Pertama sekali email yang ada dalam *folder* inbox adalah email yang belum dikategorikan. Kemudian, email-email tersebut akan melalui proses pengelasan supaya email-email tersebut akan dikategorikan mengikut setiap kategori yang telah ditetapkan. Contoh kategori email yang dibuat oleh pensyarah ialah subjek, pejabat am, pelajar dan sebagainya.

1.7 Kekangan dan Had Limitasi

Antara kekangan dan had limitasi semasa proses kajian dilakukan ialah :-

- i. Bilangan input data yang akan diuji agak terbatas.
- ii. Perisian SVM^{light} hanya digunakan untuk pengujian proses pengelasan.
- iii. Kategori email ditakrif terlebih dahulu.
- iv. Setiap input email yang digunakan perlu ditakrif terlebih dahulu dalam program yang digunakan untuk mendapatkan nilai *tfidf*
- v. Ketepatan untuk mengkategorikan email kemungkinan tidak mencapai 100%

1.8 Justifikasi dan Kepentingan

Kelebihan setelah melakukan kajian proses pengelasan email :-

- i) Dapat mengelaskan email mengikut kategori yang ditetapkan.
- ii) Projek ini dijalankan untuk mengkaji bagaimana Support Vector Machine (SVM) berdasarkan perisian SVM^{light} boleh digunakan untuk menyelesaikan masalah pengelasan email.
- iii) Pengelasan dilakukan ke atas kandungan email di dalam bahasa melayu dan bahasa inggeris

1.9 Kesimpulan

Pada masa kini, ramai pengguna menggunakan email untuk berkomunikasi dan menghantar mesej antara satu sama lain. Proses pengelasan email memudahkan pengguna membaca maklumat email yang sampai ke dalam akaun email. Kajian yang dilakukan ini akan membantu pembangun lain untuk membangunkan satu prototaip yang dapat melakukan kerja pengelasan email mengikut kategori yang ditetapkan. Adalah diharapkan kajian pengelasan email ini dapat dijadikan permulaan kepada penghasilan pakej yang lebih mantap dan meluas.