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*, 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. Effective 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 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 mengkelaskan 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 kefungsian 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 mengkelaskan 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

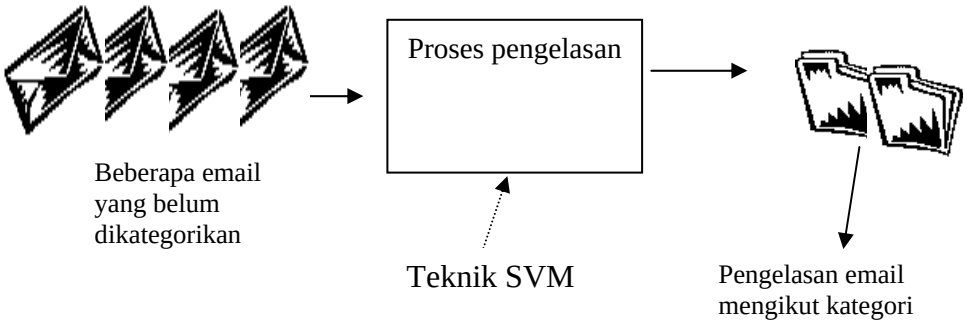
vi) Jumlah email yang digunakan semasa proses *training* (pembelajaran) adalah sebanyak 200 data email dan semasa proses *testing* (pengujian), sebanyak 100 data email

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 efisyen 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 PENGESAHAN 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**	

(TANDATANGAN PENULIS)

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

Tarikh : 18 OKTOBER 2004

Disahkan oleh

(TANDATANGAN PENYELIA)

ENCIK AHMAD FARIZ BIN ALI
Nama Penyelia

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 dimaksudan 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. Effective 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 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 mengkelaskan 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 kefungsian 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 mengkelaskan 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

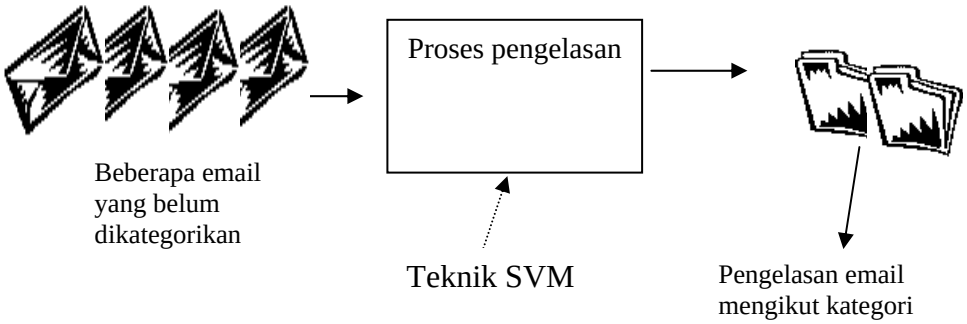
vi) Jumlah email yang digunakan semasa proses *training* (pembelajaran) adalah sebanyak 200 data email dan semasa proses *testing* (pengujian), sebanyak 100 data email

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 efisyen 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.