Jurnal Teknologi Maklumat & Multimedia **5**(2008): 91-106

Pengelasan E-mel Menggunakan Kaedah Perambat Balik

NOR AZMAN MAT ARIFF & NAZLIA OMAR

ABSTRAK

E-mel merupakan antara perkhidmatan komunikasi yang paling popular dewasa ini. Penggunaan e-mel tidak melibatkan kos yang tinggi serta pantas di dalam menyampaikan maklumat. Namun begitu, lambakan e-mel spam banyak menimbulkan masalah kepada pengguna, organisasi dan penyedia servis Internet. E-mel spam menyebabkan produktiviti kerja menurun dan kerugian dari segi penggunaan jalur lebar dan storan. Justeru itu, satu kajian telah dilakukan bagi menapis e-mel spam menggunakan rangkaian neural perambat balik. Data bagi kajian diperolehi dari e-mel peribadi penulis yang dikumpul selama 6 bulan. Perkataan yang wujud pada kandungan e-mel digunakan bagi melatih rangkaian neural. Perkataan terlebih dahulu diekstrak dari e-mel dan melalui pra proses data. Pra proses data melibatkan pembuangan kata henti, cantasan, penjanaan matriks perkataan e-mel dan umpukan pemberat terhadap perkataan. Perlaksanaan cantasan menggunakan algoritma Porter bagi perkataan bahasa Inggeris dan algoritma Fatimah bagi perkataan bahasa Malaysia. Umpukan pemberat bagi perkataan menggunakan TF-IDF dan teknik khi kuasa dua digunakan bagi memilih perkataan yang akan melatih rangkaian neural. Pemberat TF-IDF perkataan akan ditukar ke nilai 0 hingga 1 menggunakan pernormalan minimum-maksimum sebelum menjadi input kepada rangkaian neural. Kriteria pemilihan model terbaik adalah berdasarkan kepada ketepatan ramalan set latihan tertinggi bagi rangkaian neural. Hasil eksperimen dibandingkan dengan kajian lepas mendapati gabungan pemberat TF-IDF dan khi kuasa dua memberikan keputusan ramalan yang memuaskan.

Katakunci: Pengelasan e-mel spam, pra pemprosesan data, rangkaian neural

ABSTRACT

E-mail is one of the most popular communication services nowadays. E-mail usage is very cost-effective and time saving for information dissemination. However, as the spamming grows, e-mail usage has created problems to the users, organizations and the Internet service providers. Spam e-mails have resulted in lower productivity and caused a big loss to bandwidth usage and

storage capacity. Hence, a study was conducted to filter the spam e-mails using the back-propagation neural network technique. The data for the study was collected from the author's personal e-mail messages for almost about 6 months. Text from the e-mail contents are used to train the neural network components. The text are extracted using some data pre-processing mechanisms; stopword discarding, stemming, e-mail word matrices building and word weighting. The stemming process uses the Porter algorithm for English words and Fatimah algorithm for Bahasa Malaysia words. Term weighting assignment is made using TF-IDF and chi-square method to select words for neural network training set. The term weighting for the TF-IDF is transformed to values between 0 and 1 using min-max normalization as the input to the neural network. The best selection model is performed based upon the most precise prediction of the training set of the neural network. From the result of the experiments of spam e-mail filtering, it shows that the combination of TF-IDF weighting and chi-square method yields a satisfactory prediction behavior.

Keywords: Spam e-mail filtering, data preprocessing, neural network

PENGENALAN

E-mel adalah satu sistem bagi menghantar dan menerima mesej secara elektronik melalui rangkaian komputer. Ia dicipta pada tahun 1971 oleh Ray Tomlinson, lulusan sains komputer dari Institut Teknologi Massachusetts. E-mel merupakan perkhidmatan yang paling popular dan digunakan dengan meluas dewasa ini. Selain daripada mudah untuk digunakan, e-mel juga tidak melibatkan kos yang tinggi serta pantas dalam menyampaikan maklumat. Di samping pelbagai kelebihan e-mel yang dinikmati oleh pengguna, ia memberikan peluang kepada jurujual atau peniaga untuk menghantar iklan kepada pengguna e-mel lain dalam kuantiti yang besar pada kos yang sangat minima. Mesej e-mel ini diterima oleh pengguna tanpa kehendak dan kerelaan mereka. E-mel ini juga disebut sebagai e-mel sampah atau juga dikenali sebagai e-mel spam.

E-mel spam boleh disamakan dengan kertas atau surat yang dihantar oleh syarikat perniagaan bagi mempromosikan produk mereka. Perbezaannya adalah, penghantaran kertas atau surat tersebut melibatkan kos yang perlu ditanggung oleh penghantar surat bagi mengedarkannya. Berlainan pula dengan e-mel spam, kos penghantaran e-mel spam perlu ditanggung oleh penerima e-mel atau penyedia servis Internet. Menurut Spinello (1999), e-mel spam memberikan kos penghantaran kepada penerima e-mel dengan menggunakan jalur lebar Internet serta sumber sistem seperti ruang storan pelayan dan juga kos pengurusan di mana penerima e-mel terpaksa membaca dan membuang e-mel yang tidak diingini.

Pengelasan e-mel spam merupakan satu proses yang mencabar. Ini kerana tiada cara yang mudah bagi membezakan e-mel sah dan e-mel spam kecuali dengan pemerhatian pengguna. Pengelasan e-mel spam akan kelihatan lebih mudah jika penghantar e-mel spam menyatakan 'spam' pada subjek setiap e-mel yang dihantar. Namun adalah mustahil untuk berharap perkara tersebut berlaku. Oleh itu, kajian terhadap pengelasan e-mel perlu giat dijalankan dengan harapan tiada lagi e-mel spam yang dapat menembusi peti masuk pengguna e-mel. Tanpa usaha bagi mengkaji masalah ini, berkemungkinan besar pada masa akan datang pengguna akan berhenti menggunakan e-mel. Oleh itu, satu kajian bagi menangani masalah ini perlu dilakukan. Terdapat tiga objektif utama yang perlu dipenuhi, iaitu:

- Membangunkan perisian pengelasan e-mel yang berupaya mengelas e-mel berbahasa Malaysia dan Inggeris.
- Menggunakan pendekatan rangkaian neural perambat balik di dalam menyelesaikan masalah pengelasan.
- Membandingkan hasil eksperimen dengan kajian lepas.

KAJIAN LATARBELAKANG

E-mel spam yang pertama disebarkan pada rangkaian luas adalah mesej iklan daripada Digital Equipment Corporation pada tahun 1978 (Zdziarski 2005). Kajian lepas mendapati bahawa masalah pengelasan e-mel berkait rapat dengan masalah pengelasan teks. Goetschi (2004) menyatakan bahawa pengelasan e-mel spam boleh dianggap sebagai satu kes pengelasan teks. Tambahnya lagi, terdapat beberapa kaedah berasaskan pembelajaran telah digunakan di dalam pengelasan teks malahan pengelasan e-mel. Ini termasuklah *NaÔve Bayes*, *Rocchio*, *k-Nearest Neighbour* (k-NN), *Decision Trees*, *Support Vector Machines*(SVM), *Boosting Trees* dan beberapa lagi. Kajian Zaiane dan Antonie (2002) menyenaraikan kaedah *Bayesian Networks*, *Decision Trees*, rangkaian neural, *Support Vector Machines* dan *k-Nearest Neighbour* sebagai pendekatan yang sering digunakan di dalam pengelasan teks. Namun begitu, terdapat beberapa teknik lain yang tidak menggunakan pengelasan teks di dalam mengelas e-mel seperti senarai hitam, senarai putih, teknik cabar/balas dan analisis cap jari.

Terdapat beberapa kajian yang menggunakan rangkaian neural bagi mengelas e-mel spam seperti kajian Goetschi (2004), Stuart et al. (2004), Yukun et al. (2004) dan Lad (2004). Di dalam kajian Goetschi (2004), ia membandingkan 3 teknik pemilihan perkataan iaitu *DF thresholding*, khi kuasa dua dan *Information Gain*(IG). Rangkaian neural perambat balik digunakan dan 2 jenis umpukan pemberat diuji iaitu TF dan pemberat binari. Ketepatan terbaik diperolehi dari penetapan 100 perkataan input, 40 bilangan neuron dan umpukan pemberat TF dengan khi kuasa dua sebagai teknik pemilihan perkataan. Ketepatan bagi e-mel spam adalah 95.7% dan 90.7%

bagi e-mel sah. Di dalam kajian Stuart et al. (2004), rangkaian neural perambat balik digunakan di mana sebanyak 800 e-mel sah dan 854 e-mel spam digunakan sebagai data input. Terdapat 17 atribut input digunakan bagi melatih rangkaian neural yang terdiri dari atribut bukan perkataan seperti bilangan perkataan yang lebih 15 huruf, bilangan perkataan yang tidak mempunyai huruf vokal dan lain-lain lagi. Keputusan terbaik diperolehi pada penetapan 12 neuron tersembunyi dan 500 *epoch* dengan ketepatan e-mel spam adalah 92.5% dan ketepatan e-mel sah adalah 91.3%.

Kajian Yukun et al. (2004) pula menggunakan kombinasi rangkaian neural *Self Organized Feature Map*(SOFM) dan PCA bagi mengelas e-mel. SOFM merupakan jenis rangkaian neural tanpa penyeliaan manakala PCA adalah singkatan dari *Principal Components Analysis* dan merupakan satu teknik yang digunakan bagi mengecilkan dimensi vektor. Umpukan pemberat yang digunakan adalah TF-IDF. Hasil eksperimen menggunakan rangkaian neural dibandingkan dengan pengelas *Bayesian* dan didapati rangkaian neural memberikan ketepatan 87.56% berbanding dengan pengelas *Bayesian* iaitu 83.28%. Kajian Lad (2004) juga menggunakan kombinasi rangkaian neural dan PCA. Pemberat yang digunakan adalah frekuensi perkataan(TF). Teknik *Difference of Means* digunakan bagi proses pemilihan perkataan. Di samping itu, kaedah heuristik yang dinamakan sebagai 'litar pintas' di dalam kajian Lad (2004) juga digunakan bagi mengecam ciri-ciri e-mel. Hasil eksperimen yang dijalankan mendapati daripada 400 e-mel bagi set ujian, 2 e-mel spam dan 2 e-mel sah telah salah dikelaskan.

Berdasarkan kepada keempat-empat kajian, didapati rangkaian neural sangat cekap di dalam pengelasan e-mel. Semua kajian yang dibincangkan menggunakan perkataan sebagai atribut input kepada rangkaian neural kecuali kajian yang dilakukan oleh Stuart et al. (2004). Bagi kajian yang menggunakan perkataan sebagai atribut input, satu teknik pemilihan perkataan akan digunakan bagi memilih perkataan yang mempunyai kuasa diskriminasi yang kuat terhadap kelas e-mel yang diuji. Namun begitu tiada kajian yang menggabungkan skema pemberat TF-IDF dan teknik pemilihan perkataan khi kuasa dua. Pada masa kini, tiada kajian pengelasan e-mel berbahasa Malaysia dilakukan. Oleh itu, pengelasan e-mel berbahasa Malaysia yang menggunakan skema pemberat TF-IDF dan teknik pemilihan perkataan khi kuasa dua akan dilaksanakan di dalam kajian ini.

METODOLOGI KAJIAN

Terdapat 3 modul yang dibangunkan iaitu modul pra-pemprosesan data, modul latihan dan ujian serta modul pengelas e-mel. Modul pra-pemprosesan data berfungsi menghasilkan data input yang bersih bagi digunakan oleh modul latihan dan ujian. Modul latihan dan ujian akan melatih rangkaian neural mengenal corak e-mel spam dan e-mel sah berdasarkan data input

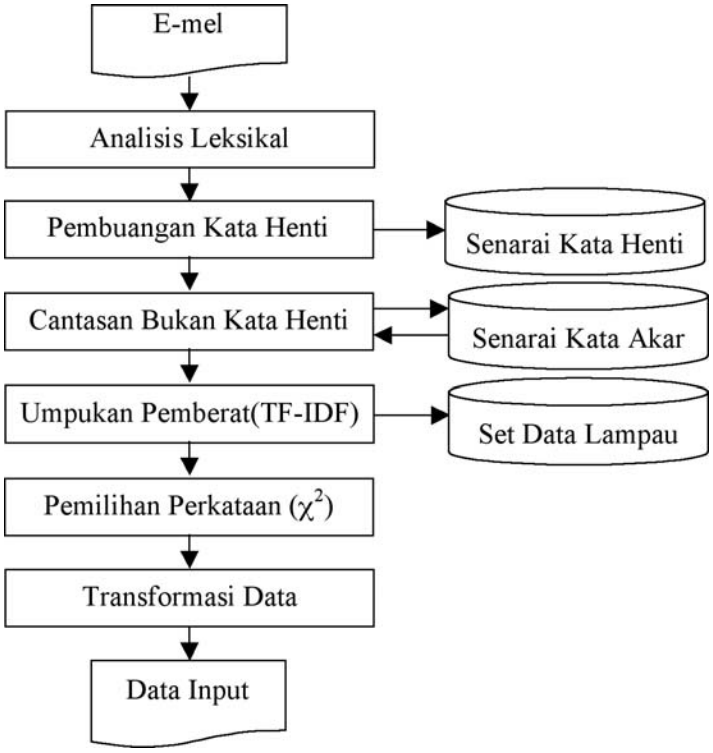
yang dihasilkan oleh modul pra-pemprosesan serta berfungsi menguji prestasi rangkaian neural berdasarkan pemberat terbaik hasil larian modul latihan. Modul pengelas e-mel pula merupakan modul perlaksanaan dan berfungsi mengelas e-mel yang dibaca dari peti masuk pengguna e-mel.

MODUL PRA-PEMPROSESAN DATA

Bagi melaksanakan pra-pemprosesan, aliran proses seperti rajah 1 digunakan. Pra-pemprosesan data merupakan antara bahagian terpenting di dalam perlaksanaan sistem berasaskan rangkaian neural. Di dalam kajian ini, pra-pemprosesan data dibahagi kepada 6 subproses iaitu analisis leksikal, pembuangan kata henti, pencantasan perkataan, pengumpukan pemberat, pemilihan perkataan dan transformasi data.

Subproses analisis leksikal melibatkan pengekstrakan isi kandungan e-mel dan dipecahkan kepada perkataan. Semua huruf bagi perkataan akan ditukarkan kepada huruf kecil. Aksara seperti delimiter asas, tanda kurungan, operator matematik dan aksara khas yang wujud di dalam perkataan juga akan dilucutkan. Bilangan huruf bagi perkataan yang diproses adalah di antara 2 hingga 15. Perkataan dengan bilangan huruf 1 dan lebih dari 15 akan direkodkan frekuensinya sahaja. Analisis leksikal berupaya mengekstrak e-mel kepada perkataan serta membuang aksara khas dan menukar semua huruf bagi perkataan kepada huruf kecil. Namun, semua perkataan yang diekstrak secara automatik akan menjadi input kepada rangkaian neural tanpa mengambilkira relevan atau tidak perkataan tersebut sebagai calon input. Bagi mengatasi masalah ini, setiap perkataan yang diekstrak dari proses analisis leksikal perlu ditapis. Semua senarai perkataan ini akan melalui proses pembuangan kata henti. Pembuangan kata henti adalah proses menyingkirkan perkataan yang kerap wujud dengan kuasa diskriminasi yang kurang seperti ‘to’, ‘a’, ‘and’, ‘it’ dan banyak lagi (Yukun et al. 2004).

Pembuangan kata henti dapat mengecilkan saiz pangkalan data yang sekaligus membolehkan hanya perkataan terpilih sahaja yang akan diproses. Namun begitu, terdapat juga perkataan yang mempunyai ejaan berlainan tetapi mempunyai makna yang sama atau secara mudahnya disebut sebagai mempunyai kata akar yang sama. Sebagai contoh, kata akar ‘makan’ mempunyai pelbagai variasi perkataan seperti ‘pemakanan’, ‘dimakan’, ‘memakan’, ‘termakan’ dan banyak lagi tetapi semua perkataan tersebut menjurus kepada satu kata akar yang sama. Jika semua perkataan ini digunakan sebagai atribut input, maka pangkalan data akan bersaiz besar. Bagi menangani masalah ini, satu proses yang dikenali sebagai cantasan perlu dilaksanakan ke atas perkataan. Dua jenis algoritma cantasan digunakan iaitu pencantas Fatimah dan pencantas Porter. Pencantas Fatimah dipilih kerana ia merupakan satu-satunya pencantas yang digunakan bagi mencantas perkataan berbahasa Malaysia. Ini kerana kebanyakan e-mel yang diterima oleh pengguna e-mel



RAJAH 1. Aliran proses modul pra-pemprosesan data

di Malaysia melibatkan perkataan berbahasa Malaysia dan Inggeris. Pencantas Porter pula digunakan bagi mencantas perkataan berbahasa Inggeris. Pencantas Porter dipilih kerana digunakan secara meluas oleh komuniti pemprosesan bahasa tabii (Sinclair dan Webber 2004; Desai 2005). Pendekatan yang digunakan bagi melaksanakan cantasan ke atas perkataan yang diekstrak dari e-mel dilakukan mengikut jujukan berikut:

- Perkataan akan dicantas menggunakan pencantas Fatimah dan Porter.
- Jika output yang terhasil dari pencantas Fatimah berbeza dengan input perkataan sebelumnya dan output yang terhasil dari pencantas Porter sama seperti input perkataan sebelumnya, maka perkataan tercantas yang dipilih adalah output yang terhasil dari pencantas Fatimah. Tetapkan status perkataan terdahulu menggunakan pencantas Fatimah.
- Jika output yang terhasil dari pencantas Fatimah sama seperti input perkataan sebelumnya dan output yang terhasil dari pencantas Porter berbeza dengan input perkataan sebelumnya, maka perkataan tercantas yang dipilih adalah output yang terhasil dari pencantas Porter. Tetapkan status perkataan terdahulu menggunakan pencantas Porter.

- Jika output yang terhasil dari pencantas Fatimah berbeza dengan input perkataan sebelumnya dan output yang terhasil dari pencantas Porter juga berbeza dengan input perkataan sebelumnya, maka pemeriksaan terhadap status perkataan terdahulu perlu dilakukan. Jika status perkataan terdahulu menunjukkan ianya menggunakan pencantas Fatimah, maka output yang terhasil dari pencantas Fatimah akan dipilih. Jika sebaliknya, maka output dari pencantas Porter akan dipilih.

Subproses pengumpukan pemberat menggunakan skema pemberat TF-IDF. TF-IDF dipilih kerana kajian lepas menunjukkan ianya merupakan skema pemberat yang sangat efisien. Ini dinyatakan di dalam kajian Robertson (2004) bahawa TF-IDF telah membuktikan kekuatan yang luar biasa dan sukar ditewaskan walaupun oleh model dan teori yang dikaji dengan teliti. TF-IDF akan memberikan nilai yang tinggi bagi perkataan yang jarang wujud dan nilai yang rendah bagi perkataan yang kerap wujud. Perkataan yang wujud pada banyak dokumen bukan pendiskriminasi yang baik dan perlu diberi pemberat yang kurang berbanding perkataan yang wujud pada sebahagian kecil dokumen (Robertson 2004). Formula bagi mengira TF-IDF perkataan adalah seperti formula (1).

$$\mathrm{TF-IDF} = \mathrm{F}_{ij} \times \log_2 \left(\frac{\mathrm{N}}{\mathrm{n}_i} \right) \tag{1}$$

di mana:

F_{ij} = Frekuensi perkataan *i* di dalam dokumen *j*.
N = Bilangan dokumen pada keseluruhan koleksi.
n_i = Bilangan dokumen di mana perkataan i wujud.

Pemilihan set perkataan bagi melatih rangkaian neural diperlukan kerana prestasi rangkaian dan kos bagi pengelasan adalah sensitif kepada saiz dan kualiti input perkataan yang digunakan untuk melatih rangkaian (Ruiz dan Srinivisan 1999). Teknik khi kuasa dua dipilih bagi melaksanakan proses pemilihan perkataan. Khi kuasa dua adalah satu teknik statistik yang sangat popular kerana ia mudah untuk dikira dan ditafsirkan (Zdziarski 2005). Bagi mengira nilai khi kuasa dua e-mel sah dan e-mel spam bagi perkataan tertentu, formula (2) digunakan:

$$\chi^2 \left(perkataan, kelas \right) = \frac{N \times \left(AD - CB \right)^2}{\left(A + C \right) \times \left(B + D \right) \times \left(A + B \right) \times \left(C + D \right)} \tag{2}$$

di mana:

A = Bilangan e-mel di dalam kelas tertentu yang mengandungi perkataan yang diproses.

B = Bilangan e-mel di dalam kelas yang lain dan mengandungi perkataan yang diproses.
C = Bilangan e-mel di dalam kelas tertentu yang tidak mengandungi perkataan yang diproses.
D = Bilangan e-mel di dalam kelas yang lain dan tidak mengandungi perkataan yang diproses.
N = Jumlah e-mel bagi set latihan.

Setiap perkataan akan mempunyai dua nilai χ^2 iaitu χ^2 sah dan χ^2 spam. Nilai χ^2 bagi satu perkataan akan dikira menggunakan formula (3) seperti di bawah:

$$\chi^2(\text{perkataan}) = P(\text{spam})\times\chi^2(\text{perkataan, spam})+P(\text{sah})\times\chi^2(\text{perkataan, sah}) \quad (3)$$

di mana:
P(spam)= Kebarangkalian wujudnya e-mel spam di dalam koleksi e-mel.
P(sah) = Kebarangkalian wujudnya e-mel sah di dalam koleksi e-mel.

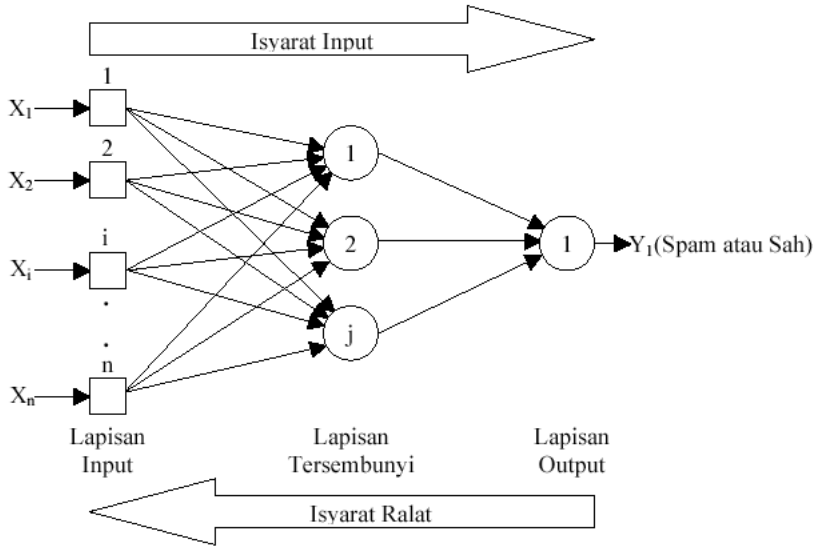
Selepas nilai χ^2 bagi setiap perkataan terhasil, perkataan akan melalui proses isihan di mana perkataan yang mempunyai nilai χ^2 tertinggi akan disusun pada kedudukan teratas. Ini bagi memberi keutamaan kepada perkataan yang mempunyai nilai χ^2 yang tinggi dipilih sebagai atribut input kepada set latihan rangkaian neural.

Proses transformasi data bertujuan menukarkan nilai pemberat terdahulu iaitu TF-IDF kepada satu nilai yang boleh diterima sebagai input kepada perambat balik. Skema transformasi data yang digunakan adalah jenis penormalan minimum-maksimum iaitu nilai TF-IDF bagi perkataan akan ditukarkan di antara julat 0 hingga 1. Formula yang digunakan adalah seperti formula (4).

$$v' = \frac{v - \min}{\max - \min}\left(new\_\max - new\_\min\right) + new\_\min \quad (4)$$

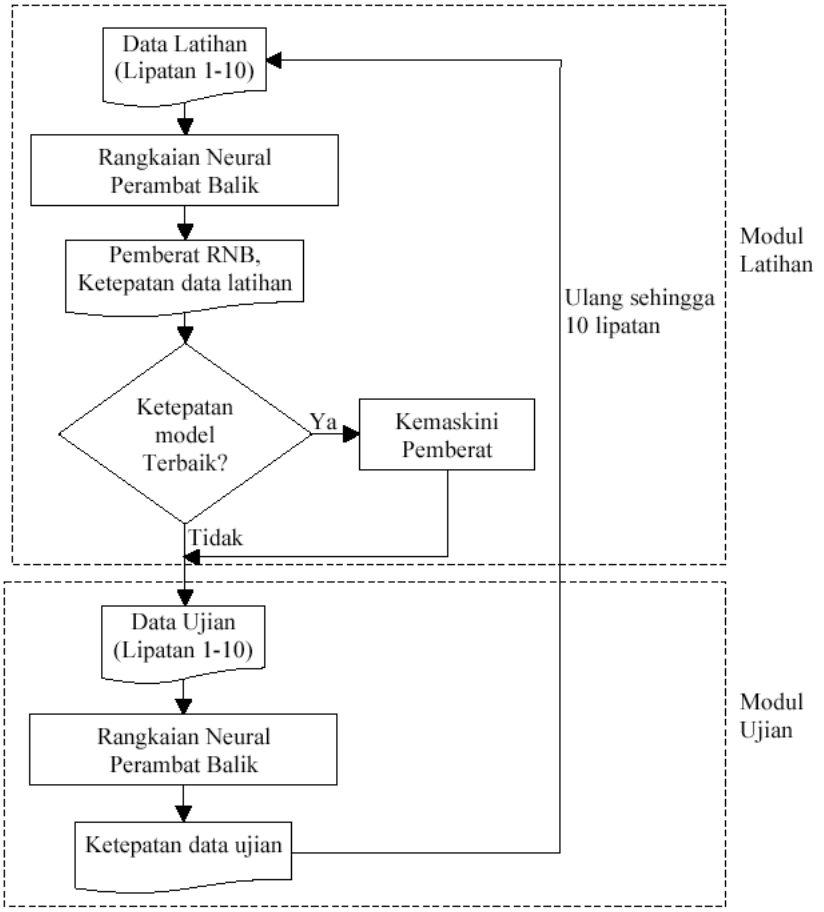
di mana:
v' = Nilai transformasi bagi perkataan.
v = Nilai TF-IDF bagi perkataan tertentu.
min = Nilai minimum TF-IDF perkataan bagi keseluruhan corak e-mel.
max =Nilai maksimum TF-IDF perkataan bagi keseluruhan corak e-mel.
new_max = Nilai 1 iaitu nilai maksimum bagi julat 0 hingga 1.
new_min = Nilai 0 iaitu nilai minimum bagi julat 0 hingga 1.

Modul latihan berfungsi menggunakan data yang telah dibersihkan sewaktu modul pra-pemprosesan bagi melatih rangkaian neural dan membina model. Di dalam kajian ini, rangkaian neural perambat balik dipilih kerana ia merupakan algoritma pembelajaran yang paling berkuasa dan digunakan dengan meluas di dalam operasi logikal kompleks, pengelasan corak dan analisis suara seperti yang dinyatakan oleh Jian et al. (2004). Di kalangan beberapa rangkaian neural buatan, rangkaian neural perambat balik amat terkenal dengan kebolehan pembelajaran yang unik (Chandren 1997). Struktur rangkaian neural perambat balik yang dibangunkan adalah seperti pada Rajah 2.



RAJAH 2. Rangkaian neural perambat balik bagi pengelasan e-mel

Pada sesi latihan, maklumat pemberat yang menghasilkan ketepatan set latihan terbaik akan disimpan. Maklumat pemberat dan templat input perkataan yang terlibat akan dikemaskini jika larian terbaru menghasilkan ketepatan set latihan terbaik yang lebih tinggi berbanding larian terdahulu. Maklumat tersebut akan digunakan bagi membina model yang akan digunakan oleh modul pengelas e-mel. Ini memastikan hanya maklumat pengelas yang terbaik sahaja disediakan kepada rangkaian neural pada modul pengelas e-mel. Bagi modul ujian pula, ia digunakan bagi mengukur prestasi model yang dihasilkan oleh modul latihan. Aliran proses modul latihan dan ujian adalah seperti Rajah 3.

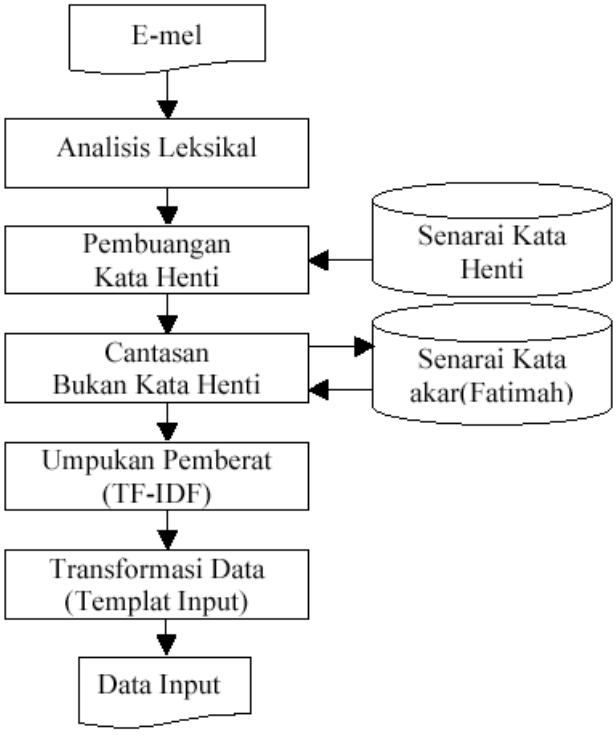


RAJAH 3. Aliran proses modul latihan dan ujian

MODUL PENGELAS E-MEL

Modul pengelas e-mel merupakan modul perlaksanaan. Di dalam modul ini, pengguna diminta untuk memasukkan 3 parameter iaitu nama pengguna, kata laluan dan nama domain pelayan e-mel. Modul pengelas e-mel akan menggunakan ketiga-tiga parameter bagi menghubungi pelayan e-mel. Proses membaca semua e-mel pengguna yang wujud pada peti masuk dilaksanakan selepas mendapat pengesahan dari pelayan e-mel. Setiap e-mel yang dibaca akan dilaksanakan pra-proses data di mana ia melibatkan subproses seperti pada modul pra-pemprosesan data kecuali subproses pemilihan perkataan. Ini kerana pemilihan perkataan telah dilaksanakan pada modul pra-pemprosesan data. Bagi subproses tranformasi data, hanya perkataan yang wujud di dalam

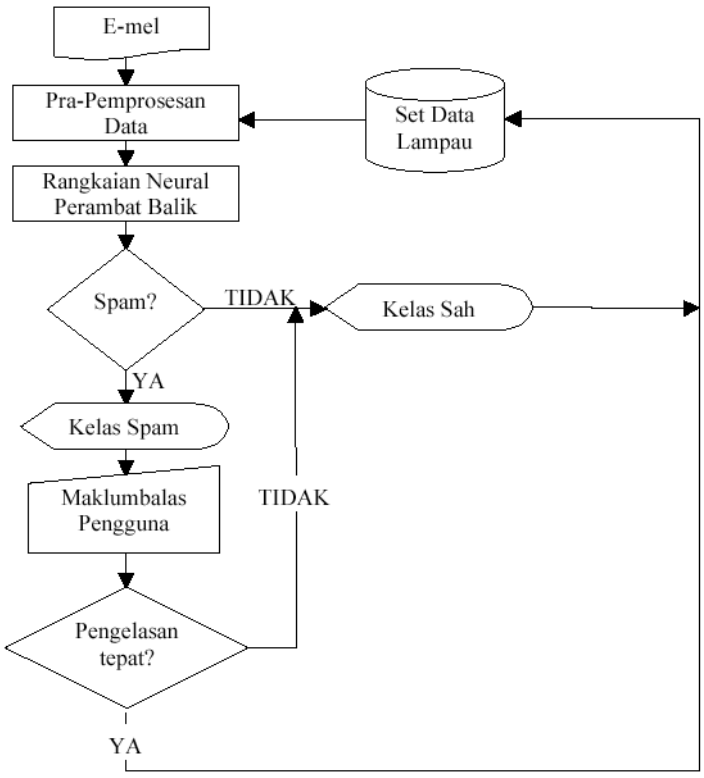
templat input perkataan sahaja terlibat dengan penukaran nilai di antara 0 hingga 1. Templat input perkataan merupakan jujukan perkataan terpilih mengikut isihan selepas proses pemilihan perkataan dilakukan pada modul latihan dan ujian terdahulu yang menghasilkan ketepatan set latihan terbaik. Aliran pra-pemprosesan data ini boleh dirujuk pada Rajah 4.



RAJAH 4. Aliran proses pra-pemprosesan data modul pengelas e-mel

Templat input perkataan dan parameter seperti pemberat lapisan input-tersembunyi dan lapisan tersembunyi-output yang menghasilkan ketepatan set latihan terbaik digunakan oleh rangkaian neural perambat balik bagi meramal status e-mel. Justifikasi pengguna diperlukan bagi menyelamatkan e-mel sah yang dikelaskan sebagai e-mel spam dengan memilih status 'bukan e-mel spam'. Begitu juga sebaliknya, e-mel spam yang diramal sebagai e-mel sah juga memerlukan bantuan pengguna bagi memilih status 'e-mel spam'. Semua maklumat perkataan bagi e-mel spam dan e-mel sah akan digunakan bagi mengemaskini set data lampau. Set data lampau merupakan senarai perkataan daripada e-mel terdahulu yang telah melalui proses pembuangan

kata henti dan pencantasan. Proses mengemaskini set data lampau ini bertujuan mengemaskini perkataan yang boleh digunakan bagi mengelas e-mel walaupun e-mel spam pada masa akan datang berubah mengikut masa. Aliran proses keseluruhan bagi modul pengelas e-mel adalah seperti pada Rajah 5.



RAJAH 5. Aliran proses keseluruhan modul pengelas e-mel

HASIL EKSPERIMEN

E-mel yang dikaji didapati daripada koleksi e-mel peribadi penulis daripada pelayan e-mel *www.kym.edu.my*. E-mel ini dikumpul selama 6 bulan dan terdapat sebanyak 500 bilangan e-mel spam dan 500 bilangan e-mel sah. Semua 500 e-mel spam yang dikumpul adalah e-mel berbahasa Inggeris manakala bagi e-mel sah pula, 250 e-mel berbahasa Malaysia dan 250 e-mel lagi berbahasa Inggeris. Ini bagi memastikan rangkaian neural mendapat pengetahuan yang pelbagai. Jumlah keseluruhan 1000 e-mel ini akan dibahagikan kepada dua set iaitu set latihan dan set ujian.

Eksperimen dijalankan dengan mengunakan 10 lipatan. Ini adalah bagi memastikan data tertabur secara seragam dan e-mel mempunyai peluang yang sama untuk dipilih sebagai set latihan dan set ujian. Bagi setiap lipatan, bilangan neuron di lapisan tersembunyi terbahagi kepada tiga iaitu bilangan neuron 10, 20 dan 30. Satu lapisan tersembunyi sahaja yang digunakan bagi keseluruhan eksperimen. Bagi setiap model, bilangan *epoch* ditetapkan kepada 1000 kali, fungsi pengaktifan adalah *hiperbolic tangent*, peratusan bagi set latihan adalah 80% dan set ujian 20%. Kriteria berhenti yang digunakan adalah berdasarkan bilangan *epoch*. Bagi setiap lipatan, terdapat 9 model yang berbeza dari segi bilangan input iaitu 100, 200 dan 300 input perkataan dan bilangan neuron 10, 20 dan 30.

Model yang mempunyai ketepatan terbaik dari set latihan bagi setiap lipatan disenaraikan seperti jadual 1. Ketepatan set ujian bagi setiap model ketepatan set latihan juga dinyatakan. Eksperimen yang dijalankan mendapati ketepatan model terbaik bagi semua lipatan tersebut menggunakan 300 bilangan input dan 10 bilangan neuron tersembunyi. Purata ketepatan model terbaik bagi semua lipatan adalah 98.78%. Ketepatan tertinggi bagi semua lipatan dihasilkan oleh model 3 pada lipatan 9 iaitu 99.12% manakala ketepatan terendah dari ketepatan terbaik semua lipatan dihasilkan oleh model 3 pada lipatan 8 iaitu 98.50%.

JADUAL 1. Model terbaik bagi 10 lipatan

Lipatan	Bil. Input	Bil. Neuron Lapisan tersembunyi	Ketepatan model	Ketepatan set ujian
1	300	10	99.00%	91.00%
2	300	10	98.62%	87.00%
3	300	10	98.87%	94.00%
4	300	10	99.00%	91.00%
5	300	10	98.75%	93.50%
6	300	10	98.62%	87.00%
7	300	10	98.62%	92.50%
8	300	10	98.50%	91.50%
9	300	10	99.12%	90.00%
10	300	10	98.62%	90.50%

Bagi melihat prestasi hasil eksperimen, perbandingan keputusan eksperimen dengan tiga kajian lepas dilakukan. Rujuk Jadual 2 bagi perbandingan hasil Stuart et al. (2004), Goetschi (2004), Lad (2004) dengan kajian semasa. Model 3 dari lipatan 9 kajian semasa iaitu ketepatan set latihan 99.12% dengan 300 perkataan input dan 10 neuron di lapisan

tersembunyi digunakan bagi perbandingan. Namun begitu, laporan bagi semua pengkaji tidak menggunakan koleksi data yang sama, Stuart et al. (2004), Goetschi (2004) serta kajian semasa menggunakan koleksi e-mel peribadi pengkaji manakala Lad (2004) menggunakan koleksi e-mel pelajar IT tahun 3 di *Indian Institute of Information Technology*, Allahabad. Jika dibandingkan ketepatan e-mel spam, hasil kajian dari Lad (2004) memberi ketepatan tertinggi berbanding ketiga-tiga kajian yang lain. Ini menunjukkan penggunaan 'litar pintas' di dalam mengenalpasti ciri e-mel spam dan pengecilan dimensi vektor menggunakan PCA yang digunakan Lad (2004) banyak membantu rangkaian neural membuat ramalan. Litar pintas adalah proses pengelasan terhadap e-mel secara heuristik tanpa melalui PCA dan rangkaian neural. Namun begitu ketepatan e-mel spam kajian semasa tidak begitu mengecewakan kerana perbezaan peratusan ketepatan yang kecil dengan tiga kajian yang lain. Hasil ketepatan e-mel spam bagi kajian semasa mengatasi hasil keputusan Stuart et al. (2004). Gabungan pemberat TF-IDF dan teknik pemilihan khi kuasa dua yang digunakan pada kajian semasa membantu rangkaian neural membuat ramalan. Kajian Stuart et al. (2004) tidak menggunakan teknik pemilihan perkataan sebaliknya menggunakan 17 atribut seperti bilangan perkataan yang lebih 15 huruf, bilangan perkataan yang tidak mempunyai huruf vokal dan lain-lain lagi. Perbezaan yang lain adalah kajian semasa menggunakan e-mel dari 2 bahasa iaitu bahasa Inggeris dan bahasa Malaysia. Oleh kerana kekurangan sampel e-mel spam berbahasa Malaysia, ini sedikit sebanyak membantu perambat balik meramal e-mel berbahasa Malaysia sebagai e-mel sah. Ini terbukti apabila hanya 6 e-mel sah sahaja yang dikelaskan sebagai e-mel spam.

Bagi ketepatan e-mel sah, hasil eksperimen Lad (2004) masih memberikan ketepatan tertinggi iaitu dengan 99.0%. Keputusan dari kajian semasa memberikan ketepatan e-mel sah terendah berbanding tiga kajian yang lain iaitu 87.61%. Ini kerana sebanyak 17.28% dari jumlah keseluruhan e-mel spam telah salah dikelaskan berbanding hanya 1% bagi Lad (2004). Semua e-mel spam yang dikelaskan sebagai e-mel sah merupakan e-mel berbahasa Inggeris dan mempunyai beberapa perkataan yang kelihatan tidak bersalah. Sebagai contoh, terdapat satu e-mel spam yang mempunyai kandungan teks berdasarkan petikan kes tragedi 11 september 2001 tetapi mempunyai pautan ke laman web yang mempromosikan jualan jam tangan. Ketepatan e-mel sah bagi Goetschi (2004) mempunyai perbezaan yang kecil dengan ketepatan yang dihasilkan Lad (2004). Hasil kajian Stuart et al. (2004) juga memberikan ketepatan e-mel sah yang baik.

Perbandingan di atas menunjukkan bahawa penyenaraian perkataan sebagai atribut input kepada perambat balik juga mampu memberikan keputusan ramalan yang baik jika digabungkan dengan teknik pemilihan perkataan dan pemberat yang tepat. Hasil dari perbandingan juga mendapati bahawa pengelasan e-mel berbahasa Malaysia adalah lebih mudah diramal oleh rangkaian neural yang dibangunkan kerana e-mel berbahasa Malaysia

JADUAL 2. Perbandingan kajian semasa dengan kajian lepas

Hasil kajian	Bil. E-mel	Bil. Input	E-mel spam		E-mel sah	
			Ketepatan	Panggilan semula	Ketepatan	Panggilan semula
Goetschi (2004)	578	100	95.7%	90.7%	98.2%	99.2%
Lad (2004)	400	6	99.0%	99.0%	99.0%	99.0%
Stuart et al.(2004)	827	17	92.45%	91.80%	91.32%	92.00%
Kajian semasa	200	300	93.10%	85.26%	87.61%	94.29%

hanya muncul pada kelas e-mel sah sahaja. Pengelasan e-mel berbahasa Inggeris menguji prestasi rangkaian neural yang sebenar kerana ianya wujud pada kedua-dua kelas e-mel iaitu e-mel spam dan e-mel sah.

KESIMPULAN

Hasil dari perlaksanaan eksperimen memberi gambaran bahawa kaedah perambat balik berupaya memberikan keputusan yang memuaskan di dalam mengelas e-mel. Pra proses data yang berkesan mempengaruhi keputusan yang dihasilkan oleh perambat balik. Penggunaan skema pemberat TF-IDF dapat memberikan perwakilan pemberat yang menyeluruh bagi koleksi e-mel berbanding teknik pemberat setempat. Menggunakan TF-IDF, perkataan yang jarang wujud pada e-mel yang lain akan diumpukkan dengan nilai yang lebih tinggi berbanding menggunakan pemberat setempat. Pemilihan perkataan input yang bertepatan banyak mempengaruhi keputusan ramalan. Tanggungjawab memilih perkataan input ini telah dilaksanakan dengan begitu baik oleh teknik khi kuasa dua. Hasil dari eksperimen menunjukkan bahawa perambat balik bertindak dengan lebih baik jika dibekalkan dengan bilangan input yang besar. Namun eksperimen yang dijalankan ditetapkan sehingga 300 perkataan input sahaja kerana melibatkan kos pengiraan yang tinggi. Bagi meningkatkan kajian pada masa akan datang, beberapa pembaikan terhadap kajian boleh dilakukan. Pemprosesan imej boleh diperkembangkan kerana terdapat e-mel spam yang mempromosikan perniagaan mereka menggunakan imej. Pemprosesan maklumat pengepala e-mel yang terperinci juga boleh dilaksanakan. Ini termasuklah maklumat alamat protokol Internet, kata nama penghantar e-mel, domain penghantar e-mel dan lain-lain lagi boleh menjadi maklumat penting di dalam pengenalpastian kelas e-mel.

RUJUKAN

Chandren, Ravie. 1997. Rangkaian Neural : Satu penerokaan dalam sistem pencapaian dokumen. Tesis Sarjana. Universiti Kebangsaan Malaysia.

Desai, N. 2005. Machine learning techniques for filtering evasive email spam. Tesis Sarjana, University of Sheffield.

Goetschi, R. 2004. Spam-filtering using artificial neural networks. Tesis Sarjana, University of Bern.

Jian, L, Guo-Yin, Z. & Guo-Chang, G. 2004. The research and implementation of intelligent intrusion detection system based on artificial neural network. *Proceedings of the Third International Conference on Machine Learning and Cybernetics*, 3178-3182.

Lad, A. 2004. SpamNet-Spam detection using PCA and neural networks. *Intelligent Information Technology*, 205-213, Heidelberg: Springer Berlin.

Robertson, S.E. 2004. Understanding inverse document frequency: on theoretical arguments for IDF. *Journal of Documentation* 60(5): 503-520.

Ruiz, M. E. & Srinivasan, P. 1999. Hierarchical neural networks for text categorization. *Proceedings of the 22nd Annual international ACM SIGIR Conference on Research and Development in information Retrieval*, 281-282.

Sinclair, G. & Webber, B. 2004. Classification from full text: A comparison of canonical sections of scientific papers. *COLING 2004 International Joint workshop on Natural Language Processing in Biomedicine and its Applications*, 69-72.

Spinello, R.A. 1999. Ethical reflections on the problem of spam. *Ethics and Information Technology* 1(3): 185-191.

Stuart, I., Cha, S. & Tappert, C. 2004. A neural network classifier for junk e-mail. *Document Analysis Systems VI,*: 442-450, Heidelberg: Springer Berlin.

Yukun, C., Xiaofeng, L. & Yunfeng, L. 2004. An e-mail filtering approach using neural network. *Advances in Neural Networks*, 688-694, Heidelberg: Springer Berlin.

Zaiane, O.R. & Antonie, M. 2002. Classifying text documents by associating terms with text categories. *Proceedings of the 13th Australasian Database Conference – Volume 5*, Darlinghurst, Australia, 215-222.

Zdziarski, J.A. 2005. *Ending Spam*. Francisco: No Starch Press, Inc.

Nor Azman Mat Ariff
Kolej Yayasan Melaka
DT 2530, Jalan SB1
Taman Seri Bayan
Gangsa, Durian Tunggal
76100 Melaka
norazmanmatariff@yahoo.com

Nazlia Omar
Jabatan Sains Komputer
Fakulti Teknologi dan Sains Maklumat
43600 UKM Bangi
Selangor
no@ftsm.ukm.my