INFORMATION SYSTEM APPLICATION, Vol. 01, No. 02
ISSN 2503-166X

Studi Perbandingan Penggabungan Metode Pemilihan Fitur dengan Metode Klasifikasi dalam Klasifikasi Teks

[1]**Genta Sahuri**
[1] President University, Jl. Ki Hajar Dewantara, Cikarang Baru – Cikarang, Bekasi 17550
E-mail: genta.sahuri@president.ac.id

***Abstract*—The main purposes of this comparative study is to obtain the best features and the method of selecting the most suitable for a particular classification method, as well as provides an overview of the performance and the accuracy of each selection features when combined with any method of classification. From the experiment it shows that for Naive Bayes classification method has the maximum degree of accuracy when combined with feature selection using Support Vector Machine. K-Nearest Neighbor classification obtains maximum accuracy when it is combined with feature selection using Information Gain and Uncertainty, with the value of *k* is 4. Furthermore, for Neural Network classifier, it looks less when it is combined with the feature selection tested since it is only produce maximum accuracy less than 50% combined with Information Gain. Moreover, Support Vector Machine resulting maximum accuracy when it is tested using Information Gain, Chi Squared, Deviation and SVM.**

Keywords— ***text classification, feature selection, classifier***

I. PENDAHULUAN

Perkembangan teknologi saat ini sangat berpengaruh pada kebiasaan manusia dalam mengkonsumsi informasi. Kecepatan dan ketepatan dalam memutuskan suatu permasalahan juga dipengaruhi oleh bagaimana media menyajikan informasi tersebut. Permasalahan yang dihadapi manusia dalam beberapa dekade kebelakang secara umum adalah tentang bagaimana mendapatkan informasi yang sesuai dengan kebutuhan. Namun dewasa ini, terjadi pergeseran dari bagaimana memperoleh informasi menjadi bagaimana menyaring informasi yang cocok dan sesuai dengan kebutuhan.

Informasi yang tersaji dari media *online* bisa berupa berita, laporan penelitian, laporan keuangan, pernyataan pada media sosial, forum diskusi dan lain sebagainya. Dengan melimpahnya data yang tersebar di internet bisa menyebabkan kesalahan penggunaan informasi dan penambahan usaha dalam memilah dan memilih dokumen yang sesuai dengan kebutuhan.

Klasifikasi teks merupakan salah satu kajian ilmu dalam bidang *Information Retrieval* dengan melakukan pengelompokan terhadap dokumen berdasarkan kategori tertentu yang sudah didefinisikan. Pemanfaatan klasifikasi teks pada tahap berikutnya bisa digunakan untuk menunjang pengambilan keputusan, pemilihan strategi pasar, perkiraan finansial, deteksi dini terhadap *email spam* atau penipuan dan lain sebagainya.

Dengan semakin berkembangnya algoritma dalam pemilihan fitur dan metode klasifikasi, muncul masalah baru, yaitu bagaimana menemukan kombinasi terbaik antara pemilihan fitur dengan metode klasifikasi. Penggunaan pemilihan fitur yang tidak pas, bisa menyebabkan performa dari metode klasifikasi menurun, sehingga tingkat akurasi yang dihasilkan rendah. Dari situ bisa ditarik kesimpulan, bahwa menemukan metode pemilihan fitur terbaik untuk digabungkan dengan metode klasifikasi sangat dibutuhkan, untuk nantinya bisa menghasilkan metode dengan tingkat akurasi yang tinggi.

Penelitian yang dilakukan adalah membandingkan metode klasifikasi dengan pemilihan fitur yang ada dengan menggunakan *dataset* yang sama untuk setiap skenario eksperimen/percobaan. Dari percobaan yang dilakukan dengan menggabungkan setiap metode pemilihan fitur dengan metode klasifikasi, akan didapatkan kesimpulan metode pemilihan fitur mana yang paling cocok dengan metode klasifikasi tertentu.

II. KAJIAN TEORI

Penelitian terkait dengan membandingkan pemilihan fitur pada beberapa metode klasifikasi telah dilakukan oleh Yiming Yang dengan membandingkan *Document Frequency Thresholding*(DF), *Information Gain* (IG), *Mutual Information* (MI), *X2 Statistic* (CHI), dan *Term Strength* (TS) [3].

Pada *Document Frequency Thresholding* (DF), Yiming Yang menghitung jumlah frekuensi dokumen untuk setiap kata yang unik dalam training *corpus* dan menghilangkan *feature space* dimana frekuensi dokumen lebih rendah dari ambang batas yang telah ditetapkan sebelumnya.

Sedangkan pada *Information Gain* dilakukan dengan mengukur jumlah *bits* dari informasi yang ditemukan untuk kategori yang diprediksi dengan mengetahui kemunculan atau ketidakmunculan kata dalam dokumen.

Untuk koleksi data yang digunakan mengambil dari Reuters-22173 dan dari data OHSUMED. Sebagai ukuran performa digunakan *recall* dan p*recision*[3].

Sebelum penerapan pemilihan fitur terhadap dokumen, terlebih dahulu dilakukan penghapusan *stop word* yang ada dalam list[3]. Selanjutnya setiap metode pemilihan fitur dievaluasi menggunakan ambang batas *term removal* yang

berbeda. Sistem SMART juga digunakan untuk proses *stemming* dan pembobotan.

Dari eksperimen yang dilakukan, diketahui bahwa pemilihan fitur menggunakan *Information Gain* dan CHI lebih efektif dalam penghilangan kata agresif tanpa kehilangan akurasi dalam klasifikasi teks. Kemampuan ambang batas dalam *document frequency* bisa dibandingkan dengan performa *Information Gain* dan CHI dengan hampir 90% penghilangan kata, sedangkan TC bisa dibandingkan dengan hamper 50-60% penghilangan kata. Sedangkan untuk *mutual information* memiliki performa paling rendah diantara pemilihan fitur yang diujikan. Hal ini terjadi karena adanya bias pendukung terhadap istilah yang jarang dan tingkat sensitifitas yang tinggi terhadap estimasi kemungkinan *error* [3].

Experimen terhadap metode *Transformed Weight Normalized Complement Naïve Bayes* (TWCNB) dilakukan dengan mengelompokan tema lagu dengan melibatkan 4 skenario. Masing masing skenario memiliki data latih yang berbeda dengan total 224 data dengan data uji tetap sebanyak 29 data. Ukuran yang digunakan dalam menentukan kualitas dari data yang terambil adalah menggunakan *precision, recall* dan *F-measure* [1].

Sedangkan Riza Ramadan [3] melakukan penelitian dengan judul Penerapan Pohon Untuk Klasifikasi Dokumen Teks Berbahasa Inggris. Pada penelitian ini digunakan 4 metode yang sudah umum dalam klasifikasi dokumen, yaitu: metode pohon atau *Iterative Dichotomiser* 3 (ID3), *Naïve Bayes, K-Nearest Neighbor* dan *Artificial Neural Network.*

Penelitian terkait lainnya dilakukan oleh Robert Neumayer, Rudolf Mayer dan Kjetil Norvag yang melakukan penggabungan metode pemilihan fitur dalam pengkategorisasian teks dimana penelitiannya menggunakan SVM dengan kombinasi pemilihan fitur: *Bi-Normal Separation, X2 statistic, Document Frequency, GSS, Information Gain, Mutual Information, Term Freq. Document Freq., Word Freq.* dan *Odd Ratio* [4].

III. METODOLOGI

Metode klasifikasi yang digunakan dalam penelitian ini adalah: *Neural Network, Naïve Bayes, K-Nearest Neighbor, Support Vector Machine.* Sedangkan pemilihan fitur yang digunakan adalah: *Information Gain, Chi Square*, dan *Gini Index*.

Adapun Evaluasi Pengukuran yang digunakan adalah *recall*, *precision*,dan *accuracy*

Dalam penelitian ini *dataset* yang akan digunakan adalah 20*NewsGroup* yang merupakan koleksi dokumen yang terdiri dari sekitar 20.000 dokumen dan dibagi dalam 20 kategori. Adapaun *dataset* yang digunakan meliputi 5 kategori, yaitu:

- rec.autos
- comp.graphics
- alt.atheism
- misc.forsale
- comp.sys.ibm.pc.hardware

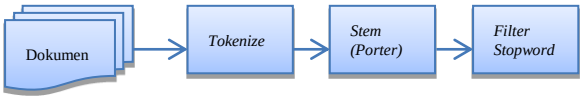
Tahapan proses penelitian yang dilakukan dengan menggunakan aplikasi *Rapidminer*.

Langkah pertama adalah dengan menentukan *dataset* yang digunakan, serta mendefinisikan label kelas untuk setiap dokumen.

Tabel 1. Dataset dan kelas label yang sudah di definisikan

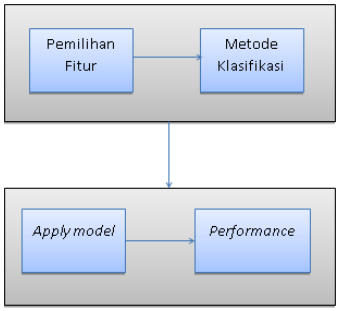
Nama kelas	Direktori
Atheism	C:\exp\alt.atheism
Comp.graphics	C:\exp\comp.graphics
Pc.hardware	C:\exp\comp.sys.ibm.pc.hardware
Forsale	C:\exp\misc.forsale
Autos	C:\exp\rec.autos

Berikutnya, dilakukan proses *Tokenizing*, *Stemming* dan *Filter Stopword* untuk mengurangi *noise* yang bisa mengurangi tingkat akurasi dari pengujian seperti yang terlihat pada Gambar 1.



Gambar 1. *Preprocessing* terhadap *dataset* yang akan diuji

Setelah dilakukan proses *preprocessing*, dilanjutkan dengan melakukan *cross validation*. Disini proses yang terjadi terbagi menjadi proses latih dan proses uji (Gambar 2).



Gambar 2. Proses Validasi

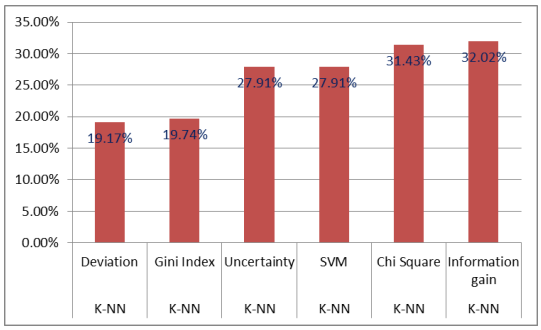
IV. HASIL DAN PEMBAHASAN

Penelitian yang dilakukan menghasilkan tingkat *accuracy*, *precision*, dan *recall* yang berbeda untuk setiap scenario ketika diuji menggunakan bobot kata yang berbeda pada metode pemilihan fitur yang dipilih. Bobot kata yang digunakan adalah pada ambang batas besar atau sama dengan 1 (>= 1.0) dan besar atau sama dengan 0.05 (>= 0.05). Untuk mengukur tingkat efektifitas dari masing masing skenario percobaan, dilakukan pengukuran dengan menggunakan *recall*, p*recision* dan *accuracy*.

Apabila bobot yang digunakan adalah >= 1, maka atribut tersebut akan dianggap dan diperhitungkan sebagai fitur, untuk selanjutnya diikutkan dalam penghitungan dengan metode

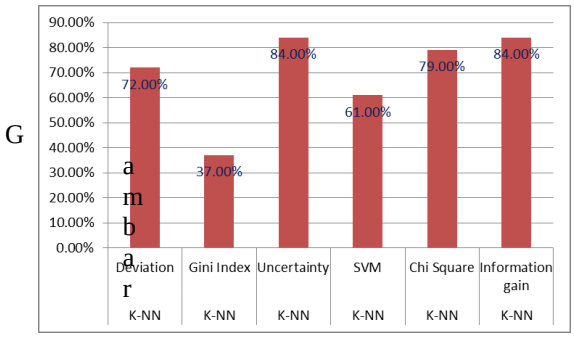
klasifikasi yang dipilih. Sebaliknya, apabila bobot yang didapat pada atribut tersebut lebih kecil dari satu, maka atribut tersebut akan diabaikan karna dianggap fitur yang tidak berpengaruh dan hanya akan membuat performa dan akurasi dari metode klasifikasi menurun.

Pada pengujian dengan menggunakan metode klasifikasi *K-Nearest neighbor* (kNN), pemilihan fitur menggunakan *Information Gain* dengan bobot kata besar atau sama dengan satu (>= 1) bekerja maksimal dibanding pemilihan fitur yang lain, yaitu dengan tingkat akurasi diangka 32.02%. Nilai *k* yang dipakai adalah 1. Ketika bobot kata diturunkan menjadi besar atau sama dengan 0.05 (>= 0.05), maka tingkat akurasi yang didapat meningkat dari sebelumnya 32.02% menjadi 79.09%. Nilai ini sedikit lebih rendah jika dibandingkan dengan metode pemilihan fitur menggunakan *Gini Index*, dimana hasil akurasi yang didapat sebesar 79.88%. Dari skenario pengujian yang dilakukan, nilai *k* yang diberikan adalah 1, sehingga algoritma *k-NN* ini hanya akan mengambil 1 dari tetangga terdekat yang memiliki tingkat kemiripan maksimal. Gambar 3. grafik perbandingan metode klasifikasi *k-NN* dengan metode pemilihan fitur yang diuji dalam penelitian ini.



Gambar 3. Grafik Perbandingan Nilai Akurasi k-NN dan Pemilihan Fitur (k=1, bobot >= 1.0)

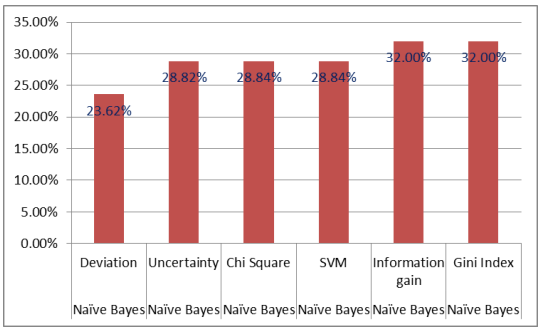
Berikutnya dilakukan pengujian ulang atas skenario di atas, tetapi dengan mengganti nilai *k* dari 1 menjadi 4 dan bobot untuk setiap fitur pada ambang batas besar atau sama dengan 0.05. Hasil pengujian ditampilkan pada Gambar 4.



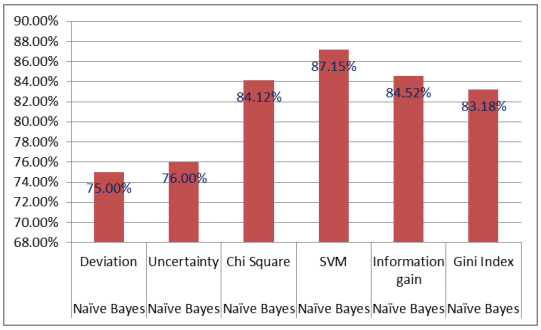
Gambar 4. Grafik Perbandingan Nilai Akurasi k-NN dan Pemilihan Fitur (*k*=4, bobot >= 0.05)

Dari skenario yang dilakukan, terlihat metode klasifikasi *k-NN* bekerja maksimal ketika dikombinasikan dengan pemilihan *fitur Information Gaindan Uncertainty*.

Selanjutnya, dalam pengujian menggunakan metode klasifikasi *Naïve Bayes* dengan bobot kata yang digunakan adalah besar sama dengan satu (>= 1), pemilihan fitur dengan menggunakan *Information Gain* dan *Gini Index* memperoleh nilai akurasi maksimal dibandingkan dengan metode pemilihan fitur lain yang diuji. Nilai akurasi yang diperoleh sebesar 32.00%.Akan tetapi ketika nilai bobot kata diturunkan menjadi besar atau sama dengan 0.05 (>= 0.05), pemilihan fitur dengan SVM menunjukkan hasil maksimal dibanding metode yang lain, yaitu dengan tingkat akurasi sebesar 87.15%. Gambar 5 dan Gambar 6 menunjukkan grafik perbandingan akurasi untuk metode klasifikasi Naïve Bayes ketika dikombinasikan dengan pemilihan fitur yang ada.

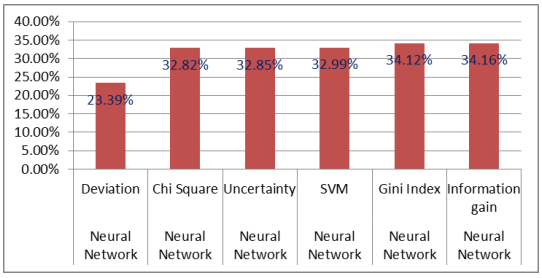


Gambar 5. Grafik Perbandingan Nilai Akurasi Naive Bayes dan Pemilihan Fitur (bobot >= 1.0)



Gambar 6. Grafik Perbandingan Nilai Akurasi Naive Bayes dan Pemilihan Fitur (bobot >= 0.05)

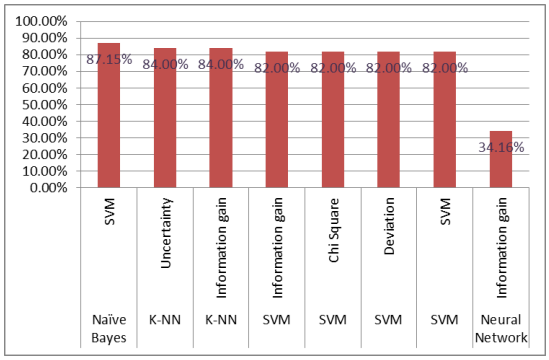
Dalam melakukan pengujian terhadap metode klasifikasi *Neural Network*, untuk pembobotan atribut dengan nilai besar atau sama dengan satu (>=1.0), diperoleh akurasi maksimal ketika menggunakan pemilihan fitur *Information Gain*, yaitu dengan nilai akurasi 34.16%, seperti yang terlihat pada Gambar 7.



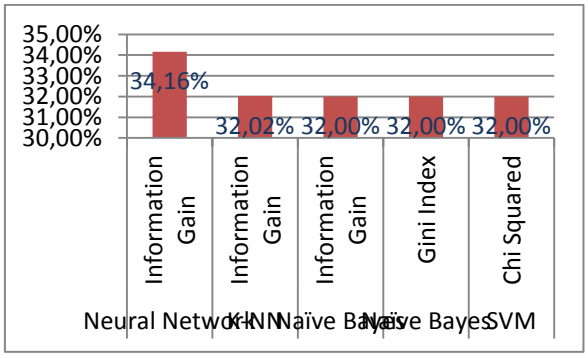
Gambar 7. Grafik Perbandingan Nilai Akurasi *Neural Network* dan Pemilihan Fitur (bobot >= 1.0)

Pengujian terakhir adalah membandingkan metode klasifikasi *Support Vector Machine*.Bobot kata yang digunakan adalah besar atau sama dengan satu (>= 1). Pemilihan fitur maksimal adalah *Chi Square* dengan tingkat akurasi 32%.

Dari semua skenario eksperimen yang sudah dilakukan, dapat disimpulkan metode klasifikasi dan metode pemilihan fitur terbaik dalam menghasilkan akurasi maksimal. Datanya akan ditampilkan pada Gambar 8, dimana metode klasifikasi *Naïve Bayes* memperoleh peringkat tertinggi dibanding metode klasifikasi lainnya ketika dikombinasikan dengan pemilihan fitur *Support Vector Machine* untuk bobot dengan nilai besar atau sama dengan 0.05. Untuk bobot dengan nilai besar atau sama dengan 1.0, metode terbaik adalah kombinasi antara *Neural Network* dengan *Information Gain* seperti yang terlihat pada Gambar 9.



Gambar 8. Grafik Perbandingan Nilai Akurasi metode klasifikasi dan pemilihan fitur terbaik untuk setiap skenario percobaan (bobot >= 0.05)



Gambar 9. Grafik Perbandingan Nilai Akurasi metode klasifikasi dan pemilihan fitur terbaik untuk setiap skenario percobaan (bobot >= 1.0)

Pada eksperimen yang dilakukan, untuk kategori *Comp.graphics* dan *Pc.hardware* secara umum dalam setiap skenario, terdapat kemiripan antara atribut pada kedua kategori tersebut, sehingga kinerja dari setiap metode klasifikasi dan pemilihan fitur jadi menurun. Berikut adalah beberapa atribut yang ditemukan pada dua kategori tersebut:

- *Graphics*
- *Scale*
- *Image*
- *Computer*

Dalam salah satu skenario pengujian, yaitu ketika menggunakan pemilihan fitur *SVM* dengan metode klasifikasi *Naïve Bayes*, ditemukan masih banyak muncul kata kata yang tidak ada arti yang masuk dalam proses perhitungan. Hal ini yang mengakibatkan penurunan nilai akurasi. Sebagai contoh, ditampilkan atribut yang tidak mengandung arti pada Tabel 1.

Pada pengujian dengan menggunakan pemilihan fitur *Chi Square* dan metode klasifikasi *Neural Network*, juga ditemukan atribut atribut yang secara harfiah tidak mengandung arti dan cenderung menjadi penyebab berkurangnya tingkat akurasi yang dihasilkan.

Tabel 1. Daftar atribut yang tidak mengandung arti dalam pegujian dengan SVM dan *Naïve Bayes.*

aa	aah	aaplai
aaa	aalborg	aardvark
aaaa	aam	aarhu
aaaaa	aamaz	aario
aaaahhh	aamir	aarnet
aaaca	aamrl	aaron
aaahhhh	aanbieden	aaronh
aaai	aangeboden	aarp
aaaread	aangezien	aatchoo
aaauuuggghhhhh	aantal	aatdb
aachen	aao	aau
aad	aaoepp	aawin
aaf	aap	ab

V. SIMPULAN

Dari penelitian yang sudah dilakukan, ketika menggunakan ambang batas bobot atribut pada level besar atau sama dengan 0.05, terlihat bahwa untuk metode klasifikasi *Naive Bayes* memiliki tingkat akurasi maksimal ketika dikombinasikan dengan pemilihan fitur menggunakan *Support Vector Machine*. Metode klasifikasi *K-Nearest Neighbor* mendapatkan akurasi maksimal ketika digabungkan dengan pemilihan fitur *Information Gain* dan *Uncertainty*, dengan nilai *k* yang digunakan adalah 4. Selanjutnya, untuk metode klasifikasi *Neural Network* menghasilkan akurasi maksimal ketika digabungkan dengan pemilihan fitur *Information Gain*. Terakhir untuk metode klasifikasi *Support Vector Machine* memperoleh nilai akurasi maksimal ketika diujikan menggunakan pemilihan fitur *Information Gain*, *Chi Squared*, *Deviation* dan *SVM*. Dari semua skenario yang dilakukan, disimpulkan bahwa kombinasi metode klasifikasi *Naive Bayes* dengan pemilihan fitur *Support Vector Machine* menempati peringkat akurasi maksimal dibanding skenario yang lain.

Skenario berikutnya adalah dengan mengganti nilai ambang batas bobot atribut menjadi besar atau sama dengan 1. Secara umum, tingkat akurasi menurun untuk setiap skenario percobaan. Terlihat pada metode klasifikasi *Naive Bayes*, akurasi maksimal tercapai ketika dikombinasikan dengan

pemilihan fitur *Information Gain* dan *Gini Index*. Selanjutnya pada metode klasifikasi *K-Nearest Neighbor*, pemilihan fitur dengan *Information Gain* memiliki tingkat akurasi maksimal dibanding yang lain. *Neural Network* memiliki nilai akurasi tertinggi ketika dikombinasikan dengan pemilihan fitur *Information Gain*. Terakhir untuk metode *Support Vector* Machine, akurasi maksimal diperoleh dengan menggunakan *Information Gain* dan *Gini Index*.

Dari semua skenario percobaan yang dilakukan, untuk pembobotan dengan nilai besar atau sama dengan 0.05 (>= 0.05), disimpulkan bahawa metode klasifikasi *Naive Bayes* dan pemilihan fitur *Support Vector Machine* menjadi skenario terbaik dengan tingkat akurasi maksimal dibanding skenario percobaan yang lain. Selanjutnya, untuk bobot dengan nilai besar atau sama dengan 1.0 (>= 1.0), skenario terbaik dengan tingkat akurasi maksimal ditemukan pada metode klasifikasi *Neural Network* dengan pemilihan fitur *Information Gain*.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada Dr. Rila Mandala, Dr. R.B Wahyu dan Dr. Tjong Wansen atas segala saran dan nasehat dan serta motivasi yang diberikan sehingga penelitian ini dapat diselesaikan.

REFERENSI

[1] Asti Tika Pratiwi, Achmad Ridok, Indriati, Klasifikasi Tema pada Lirik Lagu dengan Metode Transformed Weight-Normalized Complement Naïve Bayes (TWCNB), Repositori Jurnal Mahasiswa PTIIK UB, Vol. 3 No. 7, 2014.

[2] Ramadan, Riza "Penerapan Pohon Untuk Klasifikasi Dokumen Teks Berbahasa Inggris", Sekolah Teknik Elektro dan Informatika, Institut Teknologi Bandung.

[3] Yang, Y., Pedersen J.P.: A Comparative Study on Feature Selection in Text Categorization (pdf) Proceedings of the Fourteenth International Conference on Machine Learning (ICML'97), 412-420, 1997.

[4] Robert Neumayer, Rudolf Mayer, Kjetil Nørvåg: Combination of Feature Selection Methods for Text Categorisation, 33rd European Conference on IR Research, ECIR 2011, Dublin, Ireland, April 18-21, 2011. Proceedings, pp 763-766, ISBN: 978-3-642-20160-8, 2011.