PENERAPAN ENSEMBLE FEATURE SELECTION DAN KLASTERISASI FITUR PADA KLASIFIKASI DOKUMEN TEKS

Mediana Aryuni

Information Systems Department, School of Information Systems, Binus University
Jl. K.H. Syahdan No. 9, Palmerah, Jakarta Barat 11480
medianaD3408@gmail.com

ABSTRACT

An ensemble method is an approach where several classifiers are created from the training data which can be often more accurate than any of the single classifiers, especially if the base classifiers are accurate and different one each other. Menawhile, feature clustering can reduce feature space by joining similar words into one cluster. The objective of this research is to develop a text categorization system that employs feature clustering based on ensemble feature selection. The research methodology consists of text documents preprocessing, feature subspaces generation using the genetic algorithm-based iterative refinement, implementation of base classifiers by applying feature clustering, and classification result integration of each base classifier using both the static selection and majority voting methods. Experimental results show that the computational time consumed in classifying the dataset into 2 and 3 categories using the feature clustering method is 1.18 and 27.04 seconds faster in compared to those that do not employ the feature selection method, respectively. Also, using static selection method, the ensemble feature selection method with genetic algorithm-based iterative refinement produces 10% and 10.66% better accuracy in compared to those produced by the single classifier in classifying the dataset into 2 and 3 categories, respectively. Whilst, using the majority voting method for the same experiment, the similar ensemble method produces 10% and 12% better accuracy than those produced by the single classifier, respectively.

Keywords: *feature clustering, classification, ensemble feature selection*

ABSTRAK

Metode ensemble dapat meningkatkan akurasi klasifikasi dibandingkan dengan sistem pengklasifikasi tunggal, terutama apabila pengklasifikasi dasarnya akurat dan sangat beragam (diverse). Sedangkan klasterisasi fitur dapat mengurangi feature space dengan menggabungkan kata-kata yang mirip menjadi sebuah klaster. Tujuan penelitian yaitu membuat aplikasi klasifikasi dokumen teks menggunakan algoritma klasterisasi fitur berbasis ensemble feature selection pada sistem pengklasifikasi dasarnya. Metodologi penelitian yang dilakukan meliputi praproses dokumen teks, pembangkitan feature subspaces untuk setiap pengklasifikasi dasar menggunakan metode random subspaces yang disertai perbaikan secara iteratif dengan algoritma genetika, implementasi algoritma klasterisasi fitur pada pengklasifikasi dasar, dan integrasi hasil klasifikasi dari setiap pengklasifikasi dasar untuk memperoleh hasil klasifikasi akhir menggunakan teknik integrasi static selection dan majority voting. Penerapan algoritma klasterisasi fitur membuat pengklasifikasi dasar memakan waktu komputasi klasifikasi relatif lebih cepat dibandingkan pengklasifikasi dasar yang tidak menerapkan klasterisasi fitur. Klasifikasi dokumen teks ke dalam 2 kategori dan 3 kategori berturut-turut memberikan penghematan waktu sebesar 1.18 detik dan 27.04 detik. Hasil uji coba menunjukkan bahwa metode ensemble feature selection, disertai perbaikan fitur yang melibatkan algoritma genetika, memiliki akurasi yang relatif lebih baik dibandingkan dengan pengklasifikasi tunggal. Untuk uji coba klasifikasi ke dalam dua dan tiga kategori, teknik integrasi static selection dapat meningkatkan akurasi rata-rata sebesar 10% dan 10.66%. Sedangkan teknik integrasi majority voting untuk uji coba yang sama dapat meningkatkan akurasi rata-rata sebesar 10% dan 12%.

Kata kunci: *klasterisasi fitur, klasifikasi, ensemble feature selection*

PENDAHULUAN

Metode *Ensemble* membangun sebuah model prediktif dengan mengintegrasikan beberapa model, yang dapat digunakan untuk memperbaiki performa prediksi (Rokach, 2010). *Ensemble classifier* telah diterapkan di berbagai aplikasi sistem prediksi meliputi prediksi struktur protein (Wu, 2010), diagnosis kanker payudara (Huang, 2010), pengenalan wajah (Yu, 2009), dan klasifikasi dokumen (Bennet, 2005)(Katakis, 2010).

Peningkatan akurasi klasifikasi dokumen teks masih perlu diteliti lebih lanjut, sehingga dikembangkan sistem pengklasifikasi yang merupakan gabungan dari beberapa pengklasifikasi tunggal (*metode ensemble*). Hasil dari tiap pengklasifikasi dasar dikombinasikan untuk mengklasifikasikan data baru. Beberapa penelitian yang dilakukan oleh Bauer dan Kohavi (1999), Dietterich (2001) dan Tsymbal (2002) menunjukkan bahwa metode *ensemble* dapat meningkatkan akurasi klasifikasi dibandingkan dengan hasil sistem pengklasifikasi tunggal, apabila pengklasifikasinya akurat dan sangat beragam (*diverse*) (Dietterich, 2001). Pendekatan efektif untuk membuat sebuah *ensemble* yang akurat dan mempunyai sifat pengklasifikasi dasar yang beragam (*diverse*) adalah dengan menggunakan *ensemble feature selection* berbasis algoritma genetika (Opitz, 1999). Penelitian mengenai *ensemble feature selection* juga dilakukan oleh Huang (2010) untuk diagnosis kanker payudara. Sedangkan Yu (2009) menggunakan algoritma genetika dalam mencari *ensemble classifier* teroptimal untuk pengenalan wajah. Tsymbal (2002) mengusulkan sebuah metode *ensemble feature selection* pada pengklasifikasi *Simple Bayesian* yang diterapkan untuk mendiagnosa penyakit radang usus buntu akut, di mana metode *random subspace* dengan penggunaan fase perbaikan berbasis *hill-climbing* dapat meningkatkan akurasi dan keragaman (*diversity*) pengklasifikasi dasar. Hasil penelitian (Tsymbal, 2002) menunjukkan bahwa hasil pengklasifikasi *Simple Bayesian* dengan metode *ensemble* memiliki akurasi yang lebih tinggi dibandingkan dengan pengklasifikasi *Simple Bayesian* tunggal.

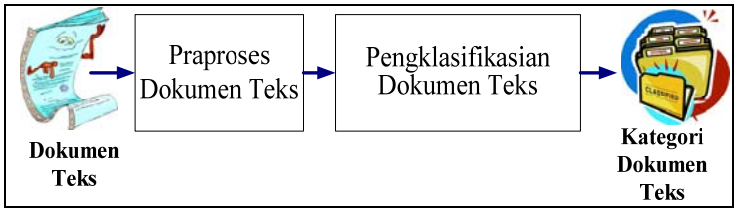
Karakteristik dokumen teks umumnya berdimensi sangat besar. Sebuah vektor dokumen dibentuk menggunakan model *bag-of-words*, di mana matriksnya cenderung bersifat *sparse*, terutama pada dokumen pelatihan yang berjumlah sedikit. Hal tersebut tidak optimal untuk algoritma klasifikasi karena model klasifikasi menjadi makin kompleks. Klasterisasi fitur dapat digunakan untuk mereduksi dimensi fitur, di mana berguna untuk mengurangi fitur yang redundan (Chen, 2004). Metode klasterisasi fitur menggabungkan kata-kata yang mirip menjadi sebuah klaster berdasarkan distribusinya. Selanjutnya klaster-klaster tersebut dipakai sebagai fitur untuk klasifikasi dokumen teks. Chen (2004) mengusulkan penghitungan similaritas antara dua klaster menggunakan informasi global yang akan memberikan keseimbangan bagi semua klaster. Hasil penelitian (Chen, 2004) menunjukkan bahwa klasifikasi dokumen teks berbasis informasi global memiliki kinerja yang lebih baik daripada sistem klasifikasi berbasis *feature selection*.

Permasalahan yang akan dibahas dalam penelitian ini, antara lain: bagaimana membangkitkan *feature subspaces* pada *ensemble feature selection* yang sesuai untuk klasifikasi dokumen teks, bagaimana membuat pengklasifikasi dasar yang menerapkan algoritma klasterisasi fitur berbasis informasi global, sehingga model klasifikasi dokumen teks pada pengklasifikasi dasar diharapkan lebih sederhana, dan bagaimana mengintegrasikan hasil klasifikasi dari tiap-tiap pengklasifikasi dasar menjadi hasil klasifikasi akhir yang diharapkan lebih akurat.

Tujuan penelitian adalah membuat aplikasi klasifikasi dokumen teks yang menerapkan algoritma klasterisasi fitur dengan menggunakan informasi global berbasis *ensemble feature selection* pada sistem pengklasifikasi dasar.

METODE

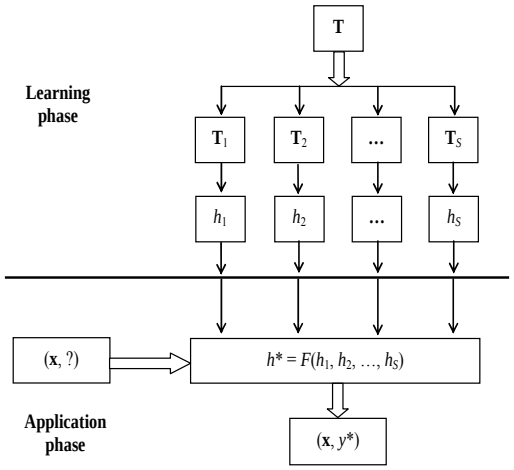
Sistem klasifikasi dokumen teks menggunakan klasterisasi fitur berbasis *ensemble feature selection* terdiri dari dua modul, antara lain: modul praposes dokumen teks dan modul klasifikasi dokumen teks. Blok diagram sistem klasifikasi dokumen teks menggunakan klasterisasi fitur berbasis *ensemble feature selection* ditunjukkan pada Gambar 1. Masukan dari sistem klasifikasi ini dibatasi berupa dokumen teks berbahasa Inggris dan hasil keluaran sistem adalah kategori dari dokumen teks tersebut.



Gambar 1 *Blok diagram* sistem klasifikasi dokumen teks menggunakan klasterisasi fitur berbasis *ensemble feature selection*

Metode Ensemble

Sebuah pengklasifikasi *ensemble* adalah himpunan beberapa pengklasifikasi, di mana keputusan dari tiap pengklasifikasi dikombinasikan dengan suatu cara (pada umumnya menggunakan *voting* dengan atau tanpa menggunakan bobot) untuk mengklasifikasikan data baru (Dietterich, 2001). Kerangka kerja dasar sebuah *ensemble* yang terdiri dari sejumlah S pengklasifikasi dasar $h_1, h_2, \ldots, h_S$ ditunjukkan pada Gambar 2.



Gambar 2 Kerangka kerja dasar sebuah *ensemble* yang terdiri dari sejumlah S pengklasifikasi dasar $h_1, h_2,.., h_S$.

Pengklasifikasi *ensemble* tersebut dibuat dengan menggunakan data pelatihan untuk membangun beberapa pengklasifikasi dasar yang kemudian digabungkan untuk memperoleh hasil klasifikasi akhir. Umumnya pada fase pembelajaran, tiap pengklasifikasi dasar h_i di dalam sebuah *ensemble* $E_l = \{h_1, \ldots, h_S\}$ dari sejumlah S pengklasifikasi dasar akan dilatih dengan menggunakan *subset* dari himpunan data pelatihan TS. Sedangkan langkah pertama pada fase aplikasi untuk mengklasifikasikan sebuah data baru x_j yaitu membangkitkan hasil klasifikasi y_i (di mana $y_i=h_i(x_j)$)

dari tiap-tiap pengklasifikasi dasar h_i dan kemudian menggunakan metode integrasi $F(y_1, \ldots, y_S)$ untuk membentuk hasil klasifikasi akhir $y^* = F(y_1, \ldots, y_S)$.

Sebuah pengklasifikasi *ensemble* akan menjadi pengklasifikasi yang efektif apabila memiliki tingkat akurasi klasifikasi yang tinggi di mana masing-masing hasil prediksinya tidak saling berkaitan (Bauer dan Kohavi, 1999). Beberapa metode untuk membuat *ensemble*, antara lain pelatihan pengklasifikasi dasar pada *subset* data pelatihan yang berbeda (misalnya *bagging* dan *boosting*), penggunaan algoritma pembelajaran yang berbeda, melakukan randomisasi, dan pelatihan pada *feature subset* yang berbeda (Dietterich, 2001).

Penerapan metode *ensemble feature selection* merupakan salah satu pendekatan yang efektif untuk membuat sebuah *ensemble* yang akurat serta pengklasifikasi dasar yang beragam (*diverse*)(Opitz, 1999), sehingga dapat menghasilkan tingkat akurasi yang lebih tinggi daripada pengklasifikasi tunggal (Dietterich, 2001).

Pengukuran tingkat keragaman div_i sebuah pengklasifikasi dasar h_i dengan keseluruhan *ensemble* $\{h_1, \ldots, h_S\}$ pada sejumlah M data pada *validation set* VS sebagai selisih rata-rata dalam klasifikasi dari semua kemungkinan pasangan pengklasifikasi dasar termasuk *hi*:

$$div_i = \frac{\sum_{j=1}^{M} \sum_{k=1, k \neq i}^{S} Dif\left(h_i(x_j), h_k(x_j)\right)}{M \cdot (S-1)} \quad (1)$$

di mana $h_i(x_j)$ menunjukkan hasil klasifikasi data x_j dari pengklasifikasi h_i, dan $Dif(h_i(x_j), h_k(x_j))$ bernilai nol jika hasil klasifikasi $h_i(x_j)$ dan $h_k(x_j)$ adalah sama dan bernilai 1 jika hasil klasifikasinya berbeda.

Pemilihan *feature subsets* menggunakan metode *random subspace* (Ho, 1998), di mana pemilihan *feature subsets* dilakukan secara acak dari *dataset* yang asli. Pengukuran kualitas tiap *feature subset* dihitung menggunakan fungsi *fitness* yang diusulkan oleh Opitz (1999). Fungsi $Fitness_i$ dari sebuah pengklasifikasi h_i yang berkaitan dengan *feature subset* FS_i adalah proporsional terhadap akurasi klasifikasi acc_i dan keragaman div_i dari pengklasifikasi h_i, di mana keduanya dihitung pada *validation set* VS:

$$Fitness_i = acc_i + \alpha \cdot div_i \quad (2)$$

di mana α adalah koefisien yang menentukan derajat pengaruh keragaman.

Algoritma *Ensemble Feature Selection* untuk *Simple Bayesian Classifier* (EFS_SBC) (Tsymbal, 2002) membangun sebuah *ensemble* dari *simple Bayesian classifiers* di dalam *random subspaces* dan menggunakan *hill climbing* pada siklus perbaikan untuk memperbaiki tingkat akurasi dan keragaman dari pengklasifikasi dasar.

Jumlah fitur untuk proses klasifikasi sangat banyak sehingga tidak memungkinkan untuk dilakukan perbaikan fitur dengan *hill-climbing search*. Pada penelitian ini akan menggunakan algoritma genetika untuk perbaikan fitur (Opitz, 1999).

Terdapat dua pendekatan utama yang diterapkan pada pembentukan metode untuk integrasi di dalam *ensembles* $F(y_1, \ldots, y_S)$, antara lain: pendekatan kombinasi, di mana pengklasifikasi dasar memperoleh hasil klasifikasi dan hasil akhir disusun dari hasil klasifikasi masing-masing pengklasifikasi dasar, dan pendekatan seleksi, di mana salah satu pengklasifikasi dasar dipilih dan hasil akhir merupakan hasil dari pengklasifikasi yang terpilih.

Klasifikasi Dokumen Teks menggunakan Klasterisasi Fitur

Klasifikasi dokumen teks (*Text Categorization*) adalah proses pengklasifikasian sebuah dokumen teks menjadi satu kategori dari beberapa kategori yang telah didefinisikan dengan menggunakan informasi dari dokumen-dokumen teks yang telah diberi label untuk proses pembelajaran (Chen, 2004).

Sistem klasifikasi dokumen teks menggunakan klasterisasi fitur berbasis informasi global terdiri dari tahap-tahap, antara lain: praproses, klasterisasi fitur (kata), dan pengklasifikasi dokumen teks berbasis klaster.

Pada penelitian ini tahap-tahap praproses yang digunakan, antara lain: analisis leksikal teks untuk menangani tanda baca, digit dan bentuk huruf (huruf kecil atau kapital), penghapusan *stopword* dengan menghapus kata-kata yang sering muncul, namun tidak dapat digunakan sebagai *index terms*, penggunaan *stemming*, yang akan menghilangkan imbuhan-imbuhan yang melekat pada sebuah kata., dan pengindeksan fitur (kata).

Algoritma 1 menunjukkan algoritma klasterisasi fitur berbasis informasi global. Pada algoritma ini ditentukan M sebagai jumlah akhir dari klaster. Langkah pertama adalah pengurutan kosakata dengan penghitungan statistik X^2 (Yang, 1997) dengan variabel klas. Kemudian klaster-klaster diinisialisasi dengan M kata teratas dari daftar kata-kata yang telah terurut. Setelah itu kata-kata yang tersisa akan dikelompokkan ke klaster-klaster tersebut.

```
Input:
W – kosakata meliputi kata-kata kandidat
M – Jumlah klaster yang diinginkan
Output:
F – Klaster pembelajaran
Klasterisasi:
  1. Urutkan kosakata dengan penghitungan
     statistik X2 dengan variabel klas.
  2. Inisialisasi M klaster sebagai
     singletons dengan M kata teratas.
  3. Ulangi sampai dengan semua kata telah
     dimasukkan ke salah satu dari M
     klaster.
  (a)Hitung similaritas antara M
     klaster (persamaan 4).
  (b)Gabungkan (merge) dua klaster yang
     paling mirip, di mana akan
     menghasilkan M-1 klaster.
  (c)Ambil kata berikutnya dari daftar
     kosakata yang telah terurut.
  (d)Buat klaster baru yang berisi kata
     yang baru.
```

Algoritma 1. Algoritma *globalCM*

Penelitian oleh McCallum (1998) menunjukkan bahwa klasifikasi *naïve bayes* dengan *multinomial event model* lebih baik daripada klasifikasi *naïve bayes* yang menggunakan *multi-variate bernoulli event model*.

Penggunaan klaster pembelajaran sebagai fitur, maka dikembangkan sebuah pengklasifikasi berbasis klaster. Vektor-vektor dokumen dibentuk dengan menggunakan model *bag-of-clusters*. Jika kata-kata terdapat pada sebuah klaster yang sama, maka kata-kata tersebut akan disajikan sebagai simbol klaster tunggal. Setelah direpresentasikan kemudian dikembangkan sebuah pengklasifikasi

berbasis fitur-fitur tersebut. Pada penelitian ini digunakan *naive Bayes* untuk mengklasifikasikan dokumen-dokumen.

Dengan diberikan sebuah dokumen d untuk klasifikasi, maka dihitung probabilitas untuk setiap kategori c sebagai berikut:

$$P(c_j|d_i;\hat{\theta}) = \frac{P(c_j|\hat{\theta})P(d_i|c_j;\hat{\theta}_j)}{P(d_i|\hat{\theta})} \approx P(d_i|c_j;\hat{\theta}) \quad (3)$$

$$P(d_i|c_j;\hat{\theta}) = P(|d_i|)|d_i|!\prod_{t=1}^{|F|}\frac{P(f_t|c_j;\theta)^{N_{it}}}{N_{it}!} \quad (4)$$

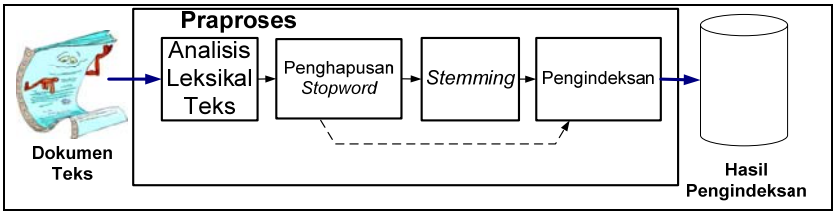
$P(c_j)$ adalah *class prior probabilities*, $|d_i|$ adalah panjang dari dokumen d_i, N_{it} adalah frekuensi dari fitur f_t (fitur-fitur adalah simbol-simbol klaster) dari dokumen d_i, F adalah fitur (klaster pembelajaran) dan |F| adalah ukuran dari F, f_t adalah fitur (klaster pembelajaran) ke-t, dan $P(f_t|c_j)$ merepresentasikan probabilitas fitur f_t merupakan kategori c_j. Probabilitas tersebut diestimasi dengan formula berikut:

$$P(f_t|c_j;\theta) = \frac{1+\sum_{i=1}^{|D|}N_{it}P(c_j|d_i)}{|F|+\sum_{s=1}^{|F|}\sum_{i=1}^{|D|}N_{is}P(c_j|d_i))} \quad (5)$$

di mana: $P(c_j|d_i) = \{0,1\}$ dan |D| = jumlah dokumen pelatihan.

Modul Praproses Dokumen Teks

Blok diagram modul praproses ditunjukkan pada Gambar 3, yang terdiri dari: modul analisis leksikal, modul penghapusan *stopword*, modul *stemming*, dan modul pengindeksan fitur (kata).



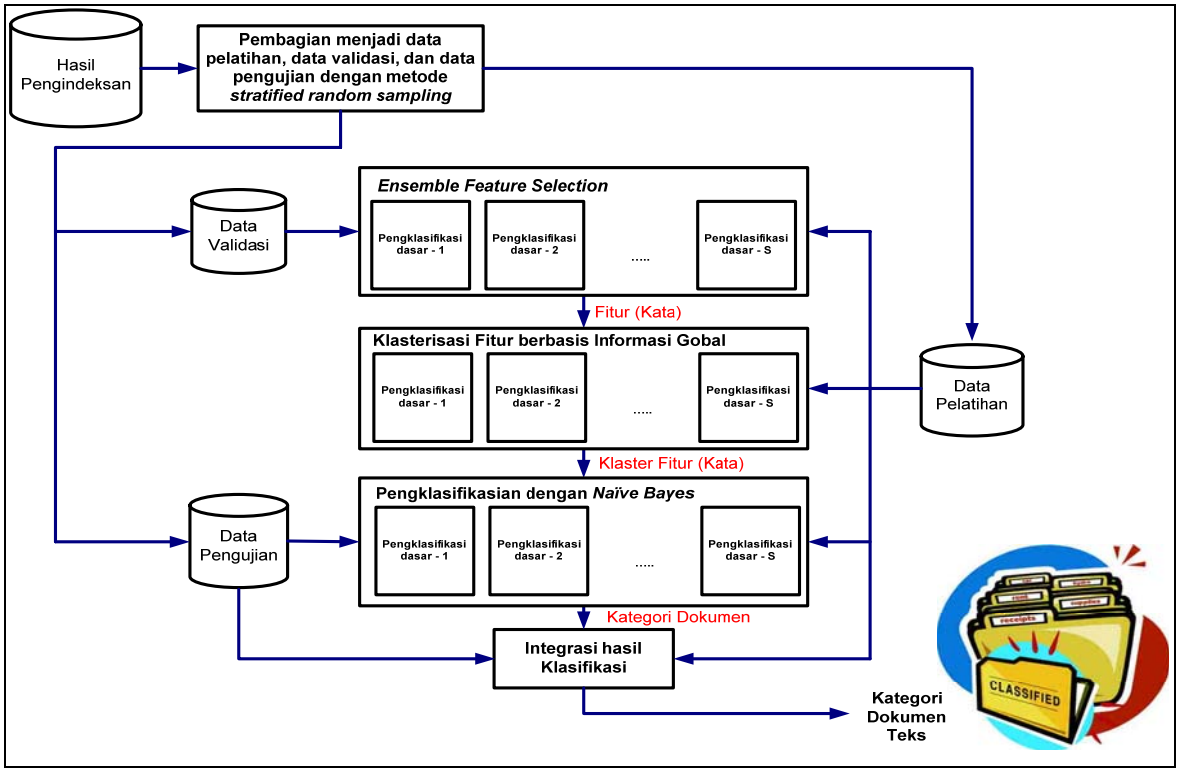
Gambar 3 Blok diagram modul praproses

Masukan modul praproses adalah dokumen teks berbahasa Inggris dan keluaran modul tersebut berupa hasil pengindeksan. Hasil pengindeksan akan digunakan untuk mendapatkan token-token yang akan menjadi input bagi sistem klasifikasi. Berdasarkan gambar 3 dapat dilihat bahwa terdapat dua jenis garis, yakni garis penuh dan garis putus-putus. Garis penuh menunjukkan modul praproses secara normal (dengan menggunakan modul *stemming*), sedangkan garis putus-putus menunjukkan modul praproses tanpa menggunakan modul *stemming*. Modul penghapusan *stopwords* dilakukan untuk mengetahui apakah sebuah kandidat kata itu termasuk *stopword* atau tidak. Jika termasuk *stopword*, maka dilanjutkan ke kandidat kata berikutnya untuk diproses. *Stopword* yang digunakan adalah *stoplist* yang disusun oleh Gerard Salton dan Chris Buckley untuk aplikasi *SMART*

Information Retrieval System di Universitas Cornell. *Stoplist* ini terdiri dari 574 kata. Sedangkan modul *stemming* yang digunakan mengacu *source code* delphi.txt pada alamat http://www.tartarus.org/~martin/PorterStemmer/ yang menggunakan algoritma *Porter*.

Modul Klasifikasi Dokumen Teks

Blok diagram modul klasifikasi dokumen teks ditunjukkan pada Gambar 4. Tahap-tahap dari klasifikasi dokumen teks, antara lain: pembagian dokumen teks menjadi data pelatihan, data validasi, dan data pengujian dengan metode *stratified random sampling*, metode *Ensemble Feature Selection* disertai perbaikan dengan algoritma gentika, klasterisasi fitur berbasis informasi global, klasifikasi dokumen teks dengan *Naive Bayes*, dan integrasi hasil klasifikasi dari masing-masing pengklasifikasi dasar menjadi hasil klasifikasi akhir. Masukan modul klasifikasi yaitu hasil pengindeksan dari modul praproses dan hasil keluaran modul ini adalah kategori dokumen teks. Masukan dari klasterisasi fitur berbasis informasi global berupa fitur (kata) dan keluaran berupa klaster fitur (kata). Klasifikasi dokumen teks dengan *Naive Bayes* menggunakan data pelatihan dengan masukan klaster fitur (kata) dan keluarannya adalah kategori dokumen teks. Integrasi hasil klasifikasi (kategori dokumen) dari masing-masing pengklasifikasi dasar menjadi hasil klasifikasi akhir (kategori dokumen).



Gambar 4 Blok diagram modul klasifikasi dokumen teks

HASIL DAN PEMBAHASAN

Uji coba dilakukan pada *dataset Newsgroup* (20NG). *Dataset* ini dikumpulkan oleh Ken Lang dan merupakan salah satu *dataset* standar yang digunakan untuk menguji algoritma klasifikasi dokumen. Kualitas hasil klasifikasi dengan berbagai variasi uji coba diukur berdasarkan *F-measure* setelah tahap integrasi hasil semua pengklasifikasi dasar. Uji coba dengan modul penghapusan *stopword* dan modul *stemming* dilakukan untuk mengklasifikasikan dokumen ke dua kategori

(*alt.atheism* dan *comp.graphics*) dan 3 kategori (*alt.atheism*, *comp.graphics*, dan *comp.os.ms-windows.misc*), di mana jumlah dokumen untuk tiap kategori yang digunakan adalah 30 dokumen sehingga terdapat 60 dokumen dengan perbandingan 1/6 data validasi, 2/3 data pelatihan, dan 1/6 data pengujian. Nilai koefisien keragaman antar pengklasifikasi α=0.25. Sedangkan variabel untuk perbaikan fitur dengan algoritma genetika, antara lain ukuran populasi=10, jumlah generasi=10. Uji coba menggunakan teknik integrasi, antara lain *static selection* dan *majority voting*.

Uji coba dilakukan untuk melakukan perbandingan akurasi dan waktu komputasi klasifikasi yakni tanpa menggunakan metode *ensemble* (pengklasifikasi tunggal) serta dengan menggunakan metode *ensemble* untuk klasifikasi dokumen teks ke dalam 2 kategori dan 3 kategori. Berdasarkan Tabel 1 ditujukkan bahwa metode *ensemble* dengan teknik integrasi *static selection* dan *majority voting* dapat meningkatkan akurasi rata-rata sebesar 10% dibandingkan dengan pengklasifikasi tunggal. Waktu klasifikasi metode *ensemble* menjadi lebih tinggi, yakni sebesar 28.097 detik seperti ditunjukkan pada Tabel 2. Berdasarkan Tabel 3 ditunjukkan bahwa metode *ensemble* dengan teknik integrasi *static selection* dan *majority voting* dapat meningkatkan rata-rata akurasi berturut-turut sebesar 10.66 % dan 12 % dibandingkan dengan pengklasifikasi tunggal. Waktu klasifikasi metode *ensemble* menjadi lebih tinggi, yakni sebesar 55.45 detik seperti ditunjukkan pada Tabel 4.

Tabel 1 Perbandingan Rata-rata Akurasi Metode Ensemble dengan Pengklasifikasi Tunggal pada Klasifikasi ke dalam 2 Kategori

Uji Coba ke-	Akurasi (%)		
	Static Selection	***Majority Voting***	**Pengklasifikasi Tunggal**
1	100	100	90.00
2	100	100	90.00
3	100	100	90.00
4	100	100	90.00
5	100	100	90.00

Tabel 2 Perbandingan Rata-rata Waktu Klasifikasi Metode Ensemble dengan Pengklasifikasi Tunggal pada Klasifikasi ke dalam 2 Kategori

Uji Coba ke-	Metode *Ensemble*				Pengklasifikasi Tunggal
	Waktu Perbaikan Fitur (detik)	**Klasterisasi Fitur (detik)**	**Klasifikasi (detik)**	**Integrasi (detik)**	**Klasifikasi (detik)**
1	938.625	9.953	42.406	0.031	15.375
2	911.609	9.390	41.157	0.031	13.297
3	952.203	10.171	43.062	0.016	16.422
4	1105.453	11.438	46.547	0.016	16.391
5	907.234	9.515	40.828	0.016	12.031
Rata-rata	963.025	10.093	**42.800**	0.022	**14.703**

Tabel 3 Perbandingan Rata-rata Akurasi Metode Ensemble dengan Pengklasifikasi Tunggal pada Klasifikasi ke dalam 3 Kategori

Uji Coba ke-	Akurasi (%)		
	Static Selection	***Majority Voting***	**Pengklasifikasi Tunggal**
1	93.33	93.33	86.67
2	86.67	93.33	80.00

3	100.00	93.33	86.67
4	93.33	93.33	86.67
5	93.33	100.00	73.33
Rata-rata	93.33	94.67	82.67

Tabel 4 Perbandingan Rata-rata Waktu Klasifikasi Metode Ensemble dengan Pengklasifikasi Tunggal pada Klasifikasi ke dalam 3 Kategori

Uji Coba ke-	**Metode *Ensemble***				**Pengklasifikasi Tunggal**
	Waktu Perbaikan Fitur (detik)	**Klasterisasi Fitur (detik)**	**Klasifikasi (detik)**	**Integrasi (detik)**	**Klasifikasi (detik)**
1	4488.046	13.078	81.187	0.016	28.265
2	4643.531	14.188	79.750	0.031	27.063
3	4764.672	14.750	83.156	0.031	23.047
4	4885.797	15.203	79.594	0.031	26.828
5	4510.625	13.516	87.234	0.031	28.469
Rata-rata	4658.534	14.147	82.184	0.028	26.734

PENUTUP

Berdasarkan hasil uji coba yang telah dilakukan dapat diambil beberapa simpulan. Pertama, aplikasi klasifikasi dokumen teks yang menerapkan algoritma klasterisasi fitur dengan menggunakan informasi global berbasis *ensemble feature selection* pada sistem pengklasifikasi dasarnya relatif mampu mengurangi waktu komputasi dari proses klasifikasi dan mampu meningkatkan akurasi klasifikasi. Selain itu, pembangkitan *feature subspaces* dengan *ensemble feature selection* yang disertai perbaikan fitur berbasis algoritma genetika tanpa menggunakan operator mutasi mampu meningkatkan akurasi klasifikasi dokumen teks untuk proses klasifikasi kedalam 2 kategori dan 3 kategori. Kedua, penerapan algoritma klasterisasi fitur berbasis informasi global sebagai metode pengurangan fitur pada vektor dokumen membuat pengklasifikasi dasar memiliki waktu komputasi klasifikasi relatif lebih cepat dibandingkan dengan pengklasifikasi dasar yang tidak menerapkan klasterisasi fitur. Hal ini ditunjukkan oleh hasil uji coba bahwa waktu komputasi yang dapat dihemat dari sebuah pengklasifikasi dasar pada klasifikasi dokumen teks kedalam 2 kategori dan 3 kategori berturut-turut adalah sebesar 1.18 detik dan 27.04 detik. Terakhir, metode *ensemble* memiliki akurasi yang relatif lebih baik dibandingkan dengan pengklasifikasi tunggal. Untuk uji coba klasifikasi kedalam 2 kategori dan 3 kategori, teknik integrasi *static selection* dapat meningkatkan akurasi rata-rata berturut-turut sebesar 10% dan 10.66%. Sedangkan teknik integrasi *majority voting* untuk uji coba yang sama dapat meningkatkan akurasi rata-rata berturut-turut sebesar 10% dan 12%.

Dalam penelitian ini, efektifitas dan efisiensi dari metode klasifikasi dokumen teks berbasis metode *ensemble feature selection* yang disertai perbaikan fitur berbasis algoritma genetika masih terbatas untuk pengujian klasifikasi dokumen teks kedalam 2 dan 3 kategori. Untuk itu diperlukan pengujian lebih lanjut untuk jumlah kategori yang lebih besar agar efektifitas dan efisiensi metode yang digunakan dapat dikaji dengan lebih mendalam. Selain itu, hasil uji coba menunjukkan bahwa secara keseluruhan waktu komputasi proses klasifikasi didominasi oleh proses perbaikan fitur. Untuk ini diperlukan penelitian lebih lanjut guna memperoleh algoritma yang mampu mengurangi komputasi dari proses perbaikan fitur.

DAFTAR PUSTAKA

Bauer, E. and Kohavi, R. (1999). An empirical comparison of voting classification algorithms: bagging, boosting, and variants. *Journal of Machine Learning*, *36* (1, 2), 105-139.

Bennet, P., Dumais, S. T. and Horvitz, E. (2005). The combination of text classifiers using reliability indicators. *Journal of Information Retrieval*. 8: 67-100

Chen, W., dkk. (2004). Automatic word clustering for text categorization using global information. *Proceedings of AIRS '04*, Beijing, China.

Dietterich, T. G. (2001). Ensemble learning methods. *Handbook of Brain Theory and Neural Networks, M.A. Arbib (ed.)* (2nd ed). Cambridge: MIT Press.

Ho, T. K. (1998). The Random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 20* (8), 832-844

Huang, M. L., Hung, Y. H. and Chen, W. Y. (2010). Neural network classifier with entropy based feature selection on breast cancer diagnosis. *Journal of Medical System*, *34*, 865–873.

Katakis, I., Tsoumakas, G. and Vlahavas, I. (2010). Tracking recurring contexts using ensemble classifiers: an application to email filtering. *Journal of Knowledge Information System*, *22*, 371-391.

McCallum, A. and Nigam, K. (1998). A Comparison of event model for naïve bayes text classification. *AAAI-98 Workshop on Learning for Text Categorization*.

Opitz, D. (1999). Feature Selection for Ensembles. *Proc. 16th Conf. on Artificial Intelligence, AAAI*, 379-384.

Rokach, L. (2010). Ensemble-based classifier. *Journal of Artificial Intelligence Rev*, *33*, 1-39.

Tsymbal, A. and Puuronen S. (2002). Ensemble Feature Selection with the Simple Bayesian Classification in Medical Diagnostics. *Proceedings of the 15 th IEEE Symposium on Computer-Based Medical Systems* (CBMS 2002).

Wu, J., Li, M., Yu, L. and Wang, C. (2010). An Ensemble Classifier of Support Vector Machines Used to Predict Protein Structural Classes by Fusing Auto Covariance and Pseudo-Amino Acid Composition. *Journal of Protein J*, *29*, 62-67.

Yang, Y. and Pedersen, J. O. (1997). A comparative study on feature selection in text categorization. In D. H. Fisher, editor, *Proceedings of ICML-97, 14th International Conference on Machine Learning*. pp. 412-420, Nashville, US. San Francisco: Morgan Kaufmann.

Yu, Z., Nam, M.Y., Sedai, S. and Rhee, P.K. (2009). Evolutionary Fusion of a Multi-Classifier System for Efficient Face Recognition. *International Journal of Control, Automation, and Systems*, *7*(1), 33-40.